CAHIER DE RECHERCHE

SEPTEMBRE 93 ■ N°48

L'ANALYSE LEXICALE :

OUTIL D'EXPLORATION DES REPRÉSENTATIONS

Réflexions illustrées par une quinzaine d'analyses

de corpus d'origines très diverses

Valérie BEAUDOUIN

Saadi LAHLOU

Avec la collaboration de Claire EVANS

Crédoc - Cahier de recherche. N° 48. Septembre 1993.

CREDOC•Bibliothèque

RÉDOC

ENTREPRISE DE RECHERCHE

Centre de Recherche pour l'Étude et l'Observation des Conditions de Vie

CRÉDOC

L'analyse lexicale : outil d'exploration des représentations

RÉFLEXIONS ILLUSTRÉES PAR UNE QUINZAINE D'ANALYSES DE CORPUS D'ORIGINES TRÈS DIVERSES

Valérie BEAUDOUIN

Saadi LAHLOU

Avec la collaboration de Claire Evans

Ce travail a bénéficié d'une subvention
du Commissariat Général du Plan
dans le cadre du programme général de recherches du CREDOC

Département Prospective de la Consommation

SEPTEMBRE 1993

142, rue du Chevaleret
75013 - PARIS

Sommaire

INTRODUCTION 1

PREMIÈRE PARTIE : AVANCÉES MÉTHODOLOGIQUES EN ANALYSE LEXICALE 11

1. - RÉPONSES À UNE QUESTION OUVERTE : INFLUENCE DU MODE DE QUESTIONNEMENT 11

1.1. Le petit déjeuner idéal : une même question, deux modes d'enquête (téléphone, auto-administré) 12

1.2. Comparaison de la richesse lexicale 15

1.3. Les différences de contenu : surface téléphonique, profondeur de l'écrit 20

2. RÉDUCTION DE LA TAILLE DU VOCABULAIRE : ANALYSE SYNTAXIQUE, RÉSEAU SÉMANTIQUE 37

2.1. La lemmatisation 38

2.2. La reconnaissance des locutions 45

2.3. La réduction sémantique 49

DEUXIÈME PARTIE : DIVERSIFICATION DES CHAMPS D'APPLICATION ET DES OUTILS DE L'ANALYSE LEXICALE 63

1. RÉFLEXION SUR LES UNITÉS D'ANALYSE : VERS UNE DÉFINITION RELATIVISTE DU TYPE D'UNITÉ 65

1.1. Le découpage alcestien : unité de contexte élémentaire et énoncé 66

1.2. Le découpage Hyperbase : unités "naturelles" de découpage 69

2. L'APPLICATION GÉNÉRALE DE LA MÉTHODE À DES DESCRIPTIONS D'OBJETS EN LANGAGE NON NATUREL 71

2.1. Le principe de la description par traits 71

2.2. Le classement des objets décrits par des traits "en désordre" 72

2.3. Le classement d'objets décrits par des séquences de traits 75

2.4. Le classement des objets complexes en transformant des descriptions codées par la concaténation des libellés des codes 78

2.5. Le travail sur les fichiers fusionnés 81

TROISIÈME PARTIE : L'INTERPRÉTATION EN ANALYSE LEXICALE 85
1. ANALYSE LEXICALE, ANALOGIE ET COMPRÉHENSION DU MONDE 87
1.1. Descriptions du monde, construction des formes et des catégories 87
1.2. Construction empirique de catégories par analyse lexicale 95
1.3. Représentation et monde 100
2. L'EXEMPLE ET LE CONCEPT : PROTOTYPES OU DESCRIPTEURS 121
2.1. Personnes âgées et prototypes 124
2.2. Idéal et prototypes 127

CONCLUSION 129

BIBLIOGRAPHIE 131

SOMMAIRE DES GRAPHIQUES 141

SOMMAIRE DES TABLEAUX 143

GLOSSAIRE 145

INTRODUCTION

Ce rapport présente les derniers développements des recherches du Département Prospective de la Consommation du CREDOC, en analyse lexicale. Avant de décrire ces travaux, il est nécessaire de donner quelques éléments sur le contexte de cette recherche : son cadre historique, ses objectifs, ses méthodes.

Les travaux du CREDOC dans le domaine de l'analyse lexicale furent initiés par Éric Brian, dans le cadre de sa première thèse (Brian, 1983, 1984, 1986), Ludovic Lebart et Monique Becue avec le développement du logiciel SPAD.Text. Ces méthodes visaient initialement à exploiter directement des réponses à des questions ouvertes de l'enquête "Aspirations et Conditions de Vie des Français", qui suit les attitudes et opinions de la population depuis 1978.

Le traitement des questions ouvertes, qui se faisait classiquement par post-codage manuel, est une opération lourde, coûteuse et surtout très dépendante de la grille de codification et des compétences de l'opérateur. Il paraissait donc intéressant de mettre en place des méthodes plus automatisées et plus fidèles.

La mise en place des premiers programmes d'analyse lexicale, qui se fait dans le cadre du logiciel SPAD, développé au CREDOC en FORTRAN IV, est à l'origine fortement marquée par l'esprit de "l'analyse des données à la française". Les outils dominants sont l'analyse factorielle et la classification hiérarchique ascendante. Dans la philosophie de ce traitement, chaque réponse en langue naturelle est considérée comme la suite, pour un individu, de modalités oui/non sur une série de variables binaires dont chacune est un mot du vocabulaire. On considère qu'il y a autant de variables V1...Vn que de mots dans l'ensemble du corpus recueilli au cours de l'enquête, et que chaque réponse est un vecteur constitué des réponses aux questions implicites : combien de fois cet enquêté a-t-il prononcé le mot V1, le mot V2, etc.

Exemple : "le beurre est dangereux", réponse donnée par l'interviewé n°1342 est traitée comme :

	(...)	"dangereux"	"est"	"grand"	(...)	"le"	"beurre"
1342	(0)	1	1	0	(0)	1	1

Ceci revient à considérer les fichiers de réponses ouvertes comme d'immenses tableaux rectangulaires, que l'on traite par les méthodes traditionnelles d'analyse des données. Cependant, un certain nombre de particularités amènent des développements techniques ou algorithmiques particuliers. La caractéristique la plus prégnante (et la plus embarrassante, nous y reviendrons) est que ces tableaux sont "hypercreux", c'est-à-dire qu'ils contiennent principalement des zéros, puisque chaque réponse n'utilise qu'une infime partie du lexique total.
Après une première période de recherches sur les caractéristiques du matériau "réponses à des questions ouvertes" (Brian, 1983), pour des raisons économiques le travail s'oriente principalement dans le perfectionnement logiciel et le traitement empirique du matériau d'enquête (voir notamment Lebart, 1984 ; Lebart et Salem, 1988 ; Haeusler, 1985, 1986, 1987) à l'aide du logiciel SPAD.Text. On se reportera à (Yvon, 1990) pour un cadrage historique plus général de cette étape.

Une nouvelle orientation est prise en 1986 avec une recherche de modélisation des représentations, comparées aux comportements (Lahlou , Betbèze et Maffre 1987a, tome III), qui marque l'émergence d'une tendance plus psychosociale dans l'analyse lexicale au CREDOC. Les textes en langue naturelle analysés (qui ne sont plus uniquement des réponses à des questions ouvertes) sont considérés comme des discours sur le monde, dont on cherche à comprendre la structure, et plus seulement comme des modalités de réponse dont on mesure la fréquence dans la population.

Sur le plan théorique, l'approche repose sur une formalisation combinatoire du monde décrit par l'observateur (Lahlou, 1990). L'idée sous-jacente est que le monde de l'observateur est un tissu continu, et multimodal ("praxéo-discursif" dirait Flament (1989), mais contenant également les perceptions et les affects) de représentations mentales. L'auteur des discours décrit une partie de ce monde dans les textes par des combinaisons de mots. Les représentations mentales d'un objet sont exprimées par les combinaisons de mots qui commutent avec le nom de l'objet, au sens paradigmatique. Les représentations sociales de l'objet sont les parties des représentations mentales largement distribuées sur une population. Nous reviendrons sur ce point de manière plus explicite dans la troisième partie de ce rapport.

Le département Prospective de la Consommation commence alors une exploration empirique systématique des techniques disponibles sur des matériaux d'origines diverses, afin de mieux comprendre la structure des représentations et leurs biais d'énonciations au travers de la langue. Cette recherche porte sur des réponses à des questions ouvertes, mais aussi des articles de journaux, des comptes-rendus d'interviews... En particulier, on s'intéresse aux

biais de recueil, et aux limites de l'analyse induites par la nature du matériau -par exemple : perte des aspects graphiques, iconiques (Lion 1990), mise en forme dans le traitement des articles de magazines ; biais de transcription manuelle du discours et d'influence de la situation d'interview (Aucouturier et al, 1991).
Parallèlement à une recherche d'amélioration des outils logiciels disponibles, en collaboration avec notamment Max Reinert (CNRS), Geneviève Teil (École des Mines, puis CNAM), et des chercheurs du Centre d'Industrie du Langage d'IBM (Isabelle Warnesson, Chantal Bedecarrax et Charles Huot), des échanges de vues sont entrepris avec des linguistes et informaticiens de diverses structures : Pierre-Yves Raccah (CNRS), Alain Bonnet, Patrick Constant, Frédéric Pigamo, (Langage Naturel SA), Christian Jacquemin (LIANA), Etienne Brunet (INALF-CNRS), Sylvie Bruxelles (CNRS), Pierre Falzon (CNAM), Luc Quoniam (CRRM), ...
Des progrès méthodologiques interviennent par la confrontation de notre approche empirique et statistique sur gros corpus représentatifs avec les travaux de nos collègues, dont l'intelligence linguistique (notamment Reinert (1983 a et b, 1986, 1987 a et b, 1990) ou la formalisation théorique (Marchotorchino et Michaud 1979, 1981 ; Warnesson 1985 ; Bedecarrax 1988 ; Marchotorchino 1989) étaient largement plus développées que celles de l'orientation CREDOC mais tournées vers des corpus ou des champs d'intérêt assez différents (stylistique, traduction automatique, analyse bibliographique...).

Sur le plan technique, des progrès sont réalisés en ce qui concerne les modalités de recueil du matériau (utilisation de méthodes projectives a-syntaxiques inspirées de l'association libre freudienne, exploitation de matériau écrit, etc.), la lemmatisation (réduction des mots à une forme canonique, utilisation de dictionnaires... où les travaux de Max Reinert nous ont apporté une contribution essentielle), la stabilité des analyses (désambiguïsation préalable du corpus, classifications croisées, combinaison des méthodes de segmentation et d'analyse en composantes multiples...), et enfin les méthodes d'interprétation des résultats (utilisation de la FRC -formalisation en relativité complète-, de l'approche par niveaux de langue, et de transferts des théories de la critique stylistique...), les méthodes de visualisation tridimensionnelle des résultats (Mac Spin), l'interfaçage avec d'autres logiciels (notamment SAS).
Des comparaisons systématiques amènent à abandonner le logiciel SPAD.Text, dont les résultats sont trop instables, au profit du logiciel ALCESTE, développé par Max Reinert, avec qui une collaboration soutenue s'engage. Le portage d'ALCESTE sous UNIX sur station SUN est réalisé au CREDOC sous la supervision de Max Reinert. L'innovation principale par rapport au logiciel SPAD.Text est celle de la stabilité des classifications, obtenue grâce à une philosophie d'approche des données différente.

L'analyse par SPAD.Text commençait par réaliser une analyse factorielle sur le tableau individus x mots, puis, sur la base de cette analyse factorielle, à effectuer une classification ascendante hiérarchique (CAH) qui classe les réponses en discours homogènes. Cette approche, qui a fait ses preuves dans l'analyse des tableaux numériques, a l'avantage de permettre une classification en ne retenant que les premiers axes factoriels, et ainsi d'éliminer le "bruit" des données. Elle donne d'excellents résultats sur les fichiers d'enquête classiques. Malheureusement, dans le cas des matrices hypercreuses typiques des fichiers de langue naturelle, les méthodes d'analyse factorielle fournissent des résultats très sensibles aux artefacts de méthode. Cela se comprend aisément si l'on sait que le nuage de points a en général la forme d'un hérisson, contenant peu de zones étendues de forte densité et de nombreux noyaux "artefactuels" provoqués par la répétition de phrases stéréotypées ou d'expressions toutes faites. Du coup, l'analyse factorielle donne des résultats peu interprétables, car les premiers axes ne rassemblent qu'une faible part de la variance, ce qui fait perdre tout son intérêt à l'approche "à la française". Les classes les plus robustes qui orientent la CAH sont justement les petits noyaux artefactuels, qui, en grossissant au cours de la classification ascendante, attirent à eux "un peu n'importe quoi", c'est-à-dire des point provenant d'amas assez éloignés, qui ont peu de liens sémantiques avec le noyau attracteur. De surcroît, ces résultats sont instables.

La méthode adoptée avec ALCESTE consiste à rechercher d'abord quels sont, dans le discours, les agrégats de formes les plus pertinents. Elle commence par une classification. Celle-ci est descendante, ce qui permet d'évacuer dès le départ les petits noyaux très typés qui provoquent des instabilités dans la constitution des classes importantes et déforment les espaces globaux lors des analyses factorielles. Une fois les classes construites, on améliore leur stabilité par des algorithmes d'échange qui permettent une meilleure redistribution des points peu typiques, les points vraiment peu typiques sont tout simplement "non classés". Les classes de faible effectif sont éliminées de l'analyse, évitant ainsi quelques artefacts. On procède ensuite à une analyse factorielle *sur la base de l'appartenance aux classes*. Cette analyse n'apporte pas d'information supplémentaire, mais permet une représentation graphique globale du corpus qui visualise les proximités relatives des classes. Elle permet en outre à l'analyste une représentation et une manipulation plus aisée des données.

Par ailleurs, des essais sont faits pour inclure des fonctionnalités nouvelles (lemmatisation "linguistique" à base d'analyse syntaxique, utilisation d'un réseau sémantique de 20 000 concepts à trois niveaux de liens).

Enfin, des recherches théoriques sur la nature de la représentation, les heuristiques de traitement, et les effets de niveaux de langue et de style ont permis, progressivement, de mettre au point la procédure de traitement actuellement la plus utilisée :

- obtenir une image de l'objet dont on cherche à connaître la représentation sociale en recueillant un corpus de phrases "à propos de cet objet" (associations mentales) ;

- reconstruire à partir de ce corpus un espace d'association *interprétable par l'analyste* dans lequel figurent les mots ou les racines utilisés dans ce corpus, et où ces mots sont d'autant plus proches les uns des autres qu'ils sont associés dans les classes tirées de l'analyse du corpus ;

- construire, par classification ou par tests statistiques, divers indicateurs permettant de décrire et mesurer l'importance des noyaux de sens repérés dans le discours ;

- caractériser les populations où chaque noyau de sens a une prévalence significativement élevée.

La recherche méthodologique et la recherche empirique sont allées de pair, chaque nouveau matériau servant à explorer des potentialités de la technique et les dernières versions des logiciels. Les derniers temps ont vu l'application de nos techniques à des corpus de plus en plus divers. Nous avons inclus dans les annexes[1], à titre d'illustration, quelques une de ces analyses tirées de rapports d'étude non publiés afin que ces expériences puissent être capitalisées.

Nous avons continué à analyser des réponses à des questions ouvertes dans le cadre de diverses enquêtes :

- **enquête Qualité** (avril 1992) : six questions ouvertes ont été posées à un échantillon de 900 personnes pour appréhender la représentation subjective de la qualité par les consommateurs :

Si je vous dis … ["produit de bonne qualité" / "produit de mauvaise qualité" / "fromage de qualité" / "voiture de qualité (ou idéale)" / "habitat de qualité (ou idéal)" / "banque de qualité (ou idéale)"]

quels sont les premiers mots qui vous viennent à l'esprit ?

[1] *L'analyse lexicale, outil d'exploration des représentations : résultats illustratifs*, Paris, Cahier de recherche n°48 bis, 1993.

Les analyses figurent dans le Cahier de recherche n°39, *Comprendre et évaluer la qualité* (CREDOC, 1992).

- **enquête sur l'effet de serre**. EDF (GRETS) nous a fourni l'ensemble des réponses à une question ouverte sur l'effet de serre. Ce corpus a été analysé à l'aide des logiciels SPAD.T et ALCESTE, ce qui a donné lieu à une analyse comparée des performances des deux logiciels. Le détail de l'étude et de la comparaison méthodologique figure en Annexe 1.

- **enquête sur l'attachement des lectrices de magazines féminins** pour le Groupe 3 Presse. Deux questions ouvertes ont été posées à mille lectrices de presse féminine pour étudier le positionnement de Femme Actuelle par rapport aux autres magazines :

> Pouvez-vous me dire en quelques mots ce que ça vous apporte de lire (nom du magazine féminin) ?
> Pour vous, un magazine féminin pour les femmes, ça serait quoi ? Qu'est-ce que vous en attendez ?

Les résultats sont pour le moment confidentiels, car c'est le groupe de presse qui a financé la réalisation de ce travail.

- **enquête Consommation** (novembre 1992). Quatre questions ouvertes ont été posées à mille personnes :

> Si je vous dis «être heureux», à quoi pensez-vous ?
> Si vous gagniez le gros lot au Loto, que feriez-vous ?
> Si je vous dis «consommateur», quels sont les mots qui vous viennent à l'esprit ?
> Si je vous dis «publicité», quels sont les mots qui vous viennent à l'esprit ?

Les analyses correspondantes figurent dans le Cahier de recherche n°46, *Où en est la consommation aujourd'hui ? : une enquête sur le consommateur français des années 90* (Lahlou, Collerie de Borely et Beaudouin, 1993).

Une cinquième question de cette enquête :

> Un petit déjeuner idéal, à quoi ça vous fait penser ?

a été posée dans une perspective méthodologique. Cette question a également été posée à mille panélistes de la SÉCODIP qui y ont répondu par écrit. L'ensemble de ce questionnaire auto-administré a été en partie analysé par les étudiants d'un groupe de travail de l'ENSAE (Beaudouin et al., en cours). En analysant les deux corpus de réponses, nous avons eu

l'occasion de mettre en évidence l'incidence du mode de questionnement. Cette comparaison méthodologique figure dans la première partie de ce rapport.

Tous ces corpus ont été analysés à l'aide du logiciel ALCESTE et complètent l'expérience acquise dans le département par l'analyse des corpus tirés des enquêtes suivantes :

- enquête **Comportements alimentaires des Français** de 1988. Quelques questions ouvertes avaient été posées à 1600 personnes. L'analyse de l'une qui portait sur le beurre :

    Si je vous dis «beurre», quels sont les premiers mots qui vous viennent à l'esprit ?

figure dans le rapport (Yvon, 1990).

Une autre portait sur le bien-manger :

    Pour vous qu'est ce que «bien manger» ?

Elle est présentée dans (Lahlou, 1992 a).

- dans le cadre de l'enquête **Aspirations et conditions de vie** du CREDOC, la question

    Si je vous dis «bien manger», à quoi pensez-vous ?

a été posée à un échantillon de 2000 personnes. L'analyse figure dans (Lahlou, 1992 a).

Mais on a surtout commencé à analyser des corpus en langage naturel ou pseudo-naturel de natures diverses :

- des entretiens semi-directifs sur l'adaptation face au risque du SIDA fournis par le CES (Centre d'Études Sociologiques) de l'Université Saint-Louis à Bruxelles ;
- un corpus d'articles du journal des Caisses d'épargne sur une période de soixante-dix ans. Cette étude a été réalisée pour le CENCEP et avait pour objectif d'analyser le référentiel de valeurs des Caisses d'épargne et son évolution. Elle figure dans l'Annexe 3.
- les outils de communication nutritionnelle de l'industrie agro-alimentaire. Cette analyse faite à la demande d'Agnès Mignonac Conseil est présentée dans l'Annexe 2.
- un corpus de décisions civiles rendues par cinq juridictions pendant une période d'un mois pour le CERCRID (Centre de recherche critique sur le droit). L'analyse figure dans (Serverin, Bruxelles, 1992). Nous y revenons dans la troisième partie de ce rapport.
- positionnement des potages et soupes à partir de l'énoncé de leurs caractéristiques techniques et de leur recette (voir Annexe 4 et (Litman, 1993)).

En outre, des travaux réalisés dans un cadre universitaire sur un recueil de Supervielle ont permis de comparer différents logiciels et de voir leur utilité pour appréhender les composantes du style, en les confrontant avec une approche stylistique classique (Beaudouin, 1993).

La technique a également été utilisée par le département "Conditions de vie et Aspirations des Français" du CREDOC, et figure dans divers rapports (Duflos, Volatier, 1991 ; Clerc, Dufour, 1992) dont une partie est confidentielle, compte tenu des financements contractuels qui les ont rendus possibles.

L'état de l'art fait que la recherche, malgré la relative maturité des techniques utilisées, reste très exploratoire : le champ qui s'ouvre est extrêmement vaste, et les acquis vont un peu dans toutes les directions. Ce rapport présente, sous la forme de chapitres distincts et relativement autonomes, certains résultats obtenus, de manière plus ou moins succincte.

Ce rapport reproduit dans son déroulement les étapes successives de l'analyse lexicale (recueil et mise en forme du matériau, définitions des individus statistiques et des variables, analyse des données, interprétation des résultats).
La première partie est consacrée au recueil des données textuelles (influence du mode de questionnement) et à l'incidence des différents modes de réduction de la taille du vocabulaire (lemmatisation-stématisation ALCESTE, lemmatisation syntaxique, réduction sémantique) sur l'analyse.
La seconde est une réflexion sur la définition des unités textuelles (équivalent des individus statistiques) et sur la sélection de traits (ou variables) pertinents pour décrire ou analyser un objet (au sens large). On montre l'efficacité des méthodes de classification descendante pour des objets décrits en langage non naturel.
Dans la dernière partie, nous proposons une théorie de l'interprétation en analyse lexicale. En appliquant cette méthode, nous avons pu identifier qu'il existait deux types de processus mentaux mis en oeuvre pour répondre à une question. Ils aboutissent à deux types de réponses : des réponses prototypiques ou extensives et des réponses conceptuelles ou compréhensives.

Comme nous faisons tout au long de ce rapport référence à ALCESTE, nous présentons ici brièvement la méthodologie ALCESTE[2].

Les analyses lexicales effectuées par ALCESTE se déroulent en trois parties.

Étape A : préparation et codification du texte initial.

Le texte est découpé en énoncés (unités de contexte initiales "UCI"). Ces unités de contexte sont de l'ordre de quelques lignes. Ces UCI sont ensuite découpées en "UCE" (unités de contexte élémentaire), de deux ou trois lignes. Pour effectuer cette découpe, ALCESTE tient compte, autant que faire se peut, de l'importance relative des marqueurs syntaxiques.

Le texte est débarrassé de tous les accents et de tous les traits d'union éventuels. ALCESTE possède un dictionnaire des locutions qui lui permet de reconnaître les "mots composés" : *à peu près, au contraire, de temps en temps...* Ceux-ci sont ensuite considérés comme étant un seul mot. Ce dictionnaire est nécessairement incomplet mais l'utilisateur a la possibilité de le modifier et d'y introduire des formes nouvelles. On peut choisir également de signaler les locutions en introduisant directement dans le texte des "_", par exemple, pain_au_chocolat.

Il est nécessaire ensuite de choisir les mots qui seront utilisés dans l'analyse ("mots pleins"). Pour ce faire, ALCESTE procède en deux étapes : premièrement, une lemmatisation, deuxièmement une reconnaissance de certains mots-outils.

Pour effectuer la lemmatisation un algorithme reconnaît, à l'aide d'un dictionnaire des racines, les mots-outils et les racines des principaux verbes irréguliers pour les réduire à leur forme infinitive. Les formes non reconnues sont ensuite traitées grâce à un algorithme particulier en se rapportant aux mots déjà lemmatisés dans le corpus et à un dictionnaire des suffixes. On a déjà vu les avantages et les inconvénients théoriques de la lemmatisation. Le choix de la lemmatisation est ici nécessaire puisque ALCESTE effectue ses analyses essentiellement à partir des fréquences et cooccurrences des mots.

Les mots pleins sont définis par opposition aux mots outils qu'ALCESTE reconnaît grâce à un dictionnaire modifiable par l'utilisateur. Sont considérés comme mots outils les articles, les pronoms, les auxiliaires *être* et *avoir*, les chiffres et certaines locutions. Les mots pleins sont regroupés dans un fichier DICO, que l'on peut modifier pour réparer d'éventuelles erreurs de lemmatisation.

Seuls les mots pleins dont la fréquence dans le texte (ou "corpus") est supérieure à trois, sont actifs. Toutes les autres formes sont des variables supplémentaires. L'utilisateur peut imposer que certaines formes de fréquence supérieure à trois soient considérées comme variables supplémentaires, en modifiant le fichier DICB qui regroupe toutes les formes lemmatisées ainsi que leur nature. Ceci peut être particulièrement utile dans le cas d'analyses de réponses à des questions ouvertes, afin d'éviter les phénomènes d'écholalie (répétition des termes de la question). On peut inversement rendre actif un mot plein de fréquence inférieure à trois, en le rattachant à un mot lemmatisé actif.

Étape B : classification descendante hiérarchique

On obtient à la fin de l'étape A un dictionnaire des formes contenues dans le texte qui permet de construire un tableau disjonctif à double entrée croisant en ligne les unités de contexte, et en colonne les formes réduites. Les cases de ce tableau valent 1 ou 0, selon que la forme apparaît dans l'UCE ou non. En règle générale, ce tableau est grand et très creux. ALCESTE utilise une méthode de classification descendante hiérarchique qui est particulièrement adaptée pour traiter des tableaux creux.

La première classe analysée contient toutes les unités de contexte. A chaque étape on cherche la meilleure bipartition de la plus grande des classes restantes. L'algorithme permet de conserver 12 classes. Le logiciel procède en trois étapes :

- il cherche le premier facteur de l'AFC du tableau creux,
- il cherche l'hyperplan perpendiculaire au premier axe maximisant l'inertie interclasses des deux nuages,
- il choisit la plus grande des classes restantes, puis réitère.

ALCESTE ne classe pas toutes les UCE : il crée une classe résiduelle d'UCE qui n'appartiennent à aucune classe.

Le choix de la classification descendante entraîne certains problèmes, en particulier la création de classes de petit effectif. Ces classes sont souvent créées par la présence de mots de faible fréquence (cependant supérieure à trois). Elles sont séparées très tôt du reste du corpus de sorte que l'on ne peut pas les éliminer en jouant sur le nombre de classes retenues (inférieur à 12). Mettre en variables supplémentaires les mots à l'origine de la création de ces classes parasites n'est pas très satisfaisant. En effet, souvent d'autres mots émergent alors et renouvellent le phénomène.

[2] Analyse Lexicale par Contexte Etabli à partir d'une Segmentation du Texte en Enoncés. Cette méthodologie a été mise au point par Max Reinert (CNRS Toulouse).

Étape C : aide à l'interprétation des classes

Cette étape crée un certain nombre de fichiers parmi lesquels les plus utiles sont :

- la liste du vocabulaire spécifique de chaque classe (profil). Sont retenus les mots lemmatisés satisfaisant un critère d'appartenance à la classe (chi^2 à un degré de liberté supérieur à 2,7). Pour chaque forme, sont fournis : son ordre d'apparition dans le dictionnaire d'ALCESTE, le nombre d'UCI de la classe et de l'ensemble du corpus qui la contiennent, son chi^2 d'appartenance à la classe, son code d'identification.
- la liste du vocabulaire non spécifique pour chaque classe (antiprofil).
- la liste des réponses les plus caractéristiques pour chaque classe. Elles sont classées par ordre de chi^2 d'association à la classe décroissant.
- l'arbre d'agrégation qui montre dans quel ordre les classes se séparent.

(Beaudouin et al. 1993)

PREMIÈRE PARTIE : AVANCÉES MÉTHODOLOGIQUES EN ANALYSE LEXICALE

Cette première partie méthodologique et technique présente :
- l'influence du mode de questionnement sur le matériau recueilli ;
- les différents modes de réduction de la taille du vocabulaire que nous avons expérimentés (utilisation de différents types d'outils linguistiques) et leur incidence respective sur l'analyse lexicale.

1. - RÉPONSES À UNE QUESTION OUVERTE : INFLUENCE DU MODE DE QUESTIONNEMENT

> *Ce travail a pu être réalisé grâce à une collaboration étroite entre la SECODIP et le CREDOC en vue de faire progresser les méthodes d'enquête. Après un premier test des questions effectué par la SECODIP auprès de 80 panélistes, le corpus de réponses auto-administrées, analysé ici, a été recueilli par la SECODIP auprès de mille de ses panélistes. Le corpus de réponses téléphoniques a été obtenu dans le cadre de l'enquête CREDOC sur la consommation auprès d'un échantillon de mille personnes en novembre 1992.*

Grâce aux méthodes de statistique lexicale, l'analyse des réponses à une question ouverte est beaucoup plus rapide et facile qu'autrefois. On a donc tendance à multiplier les questions ouvertes dans les enquêtes. Or, il semble que les différentes techniques de questionnement sont plus ou moins bien adaptées aux questions ouvertes.

Il existe actuellement plusieurs modes d'interrogation :

- l'entretien en face à face : l'enquêteur note ou saisit sur clavier les réponses de l'enquêté.
- l'interrogation par téléphone : l'enquêteur note ou tape les réponses de son interlocuteur.

Dans ces deux premiers cas, la réponse est orale puis retranscrite.

- le questionnaire auto-administré : l'enquêté se charge lui-même de remplir le questionnaire sur support papier, sur Minitel ou sur micro-ordinateur. Cette fois la réponse est une réponse écrite.

1.1. Le petit déjeuner idéal : une même question, deux modes d'enquête (téléphone, auto-administré)

Nous avons pu mettre en place les conditions méthodologiques nécessaires à une étude comparative entre deux de ces modes de questionnement : l'enquête téléphonique et le questionnaire auto-administré. Une même question a été posée à deux échantillons de même taille (mille personnes environ) :

Un petit déjeuner idéal, à quoi ça vous fait penser ?

Le premier échantillon (1 001 personnes) a été interrogé par téléphone dans le cadre de l'enquête CREDOC sur la consommation réalisée en novembre et décembre 1992 (Lahlou, Collerie de Borely et Beaudouin, 1993). Les questions qui entouraient la question ouverte portaient sur la consommation mais pas sur le petit déjeuner.

Le second échantillon (1005 personnes) est constitué par les panélistes de la SÉCODIP qui ont répondu par écrit (enquête postale) à une série de questions sur le petit déjeuner (enquête postale).

1. Un petit déjeuner idéal, à quoi ça vous fait penser ?
.....
.....
.....
.....

2. Décrivez votre petit déjeuner habituel
- en semaine
ambiance : .....
.....
lieu : .....
heure : .....
produits consommés : .....
.....
- en week-end, en vacances
ambiance : .....
.....
lieu : .....
heure : .....
produits consommés : .....
.....

Notons que la formulation de la question était tout à fait identique pour les deux échantillons.

Mais il y avait deux critères de variation :

- le mode d'interrogation (téléphone, auto-administré) ;
- le contexte de l'enquête. Dans le cadre de l'enquête téléphonique, les questions qui entouraient la question ouverte ne portaient pas sur le petit déjeuner et ne pouvaient donc influencer les personnes interrogées. En revanche, comme l'ensemble du questionnaire auto-administré était centré sur le petit déjeuner, sa lecture a pu avoir une influence sur le contenu des réponses.

En particulier, on pourrait penser que l'enquêté, lisant les questions suivantes, se sentirait obligé de répondre à chaque rubrique (lieu, heure, produits, ...) même pour la question 1. Ce n'est pas ce qui se produit, comme on peut le voir à la lecture des réponses, souvent centrées sur une seule idée.
Par ailleurs, des résultats identiques avaient été obtenus dans la pré-enquête auto-administrée, où seule figurait la question 1 :

Un petit déjeuner idéal, à quoi ça vous fait penser ?

Les questions sur la description fine des produits, heure et lieu ont précisément été rajoutées pour obtenir systématiquement des précisions concrètes qui manquaient souvent et éviter de se retrouver avec une classe "avec produits" et une classe "sans produits", pour comparer contenu matériel et contexte dans le cas des petits déjeuners habituels.

Néanmoins, compte tenu de ces différences de protocole, il faut rester prudent en interprétant les différences dans la nature des réponses obtenues.

Un autre élément aurait pu mettre en péril la validité de la comparaison : les panélistes sont exclusivement des femmes issues de ménages représentatifs de la population française alors que l'échantillon de l'enquête téléphonique est composé d'hommes et de femmes. Nous avons donc été amenés à exclure les hommes de l'enquête téléphonique pour que la comparaison soit valable.

1.2. Comparaison de la richesse lexicale ; impact de la réduction (lemmatisation, seuil de fréquence)

Avant d'entrer dans des comparaisons plus fines (en excluant les hommes de l'échantillon interrogé par téléphone), nous avons cherché à caractériser par quelques indicateurs lexicaux les deux corpus de réponses.

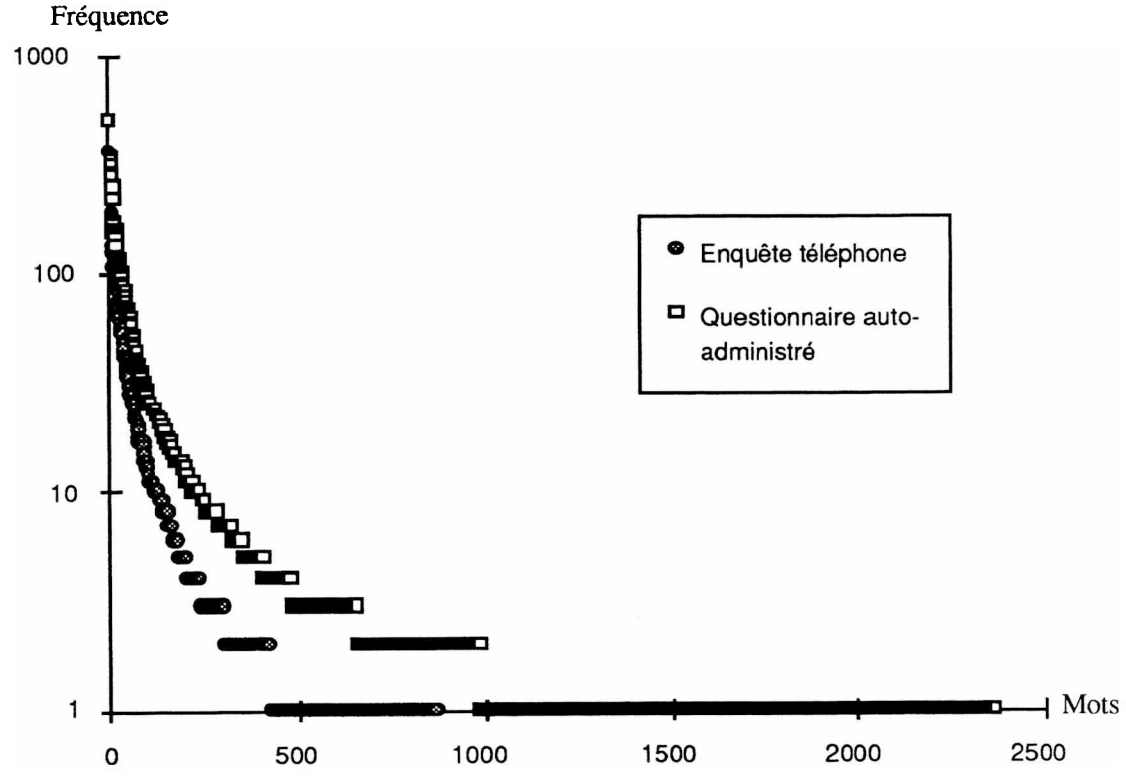
1.2.1. Les réponses écrites sont plus longues et plus variées que les réponses orales

Le nombre moyen de mots par réponse est nettement plus élevé par écrit que par téléphone. En effet , en auto-administré, une réponse est en moyenne constituée de près de 19 mots, alors que par téléphone, le nombre moyen est de 15. On a donc en moyenne trois ou quatre mots de plus par réponse dans un questionnaire écrit.

La taille du vocabulaire (le nombre de mots distincts employés) est beaucoup plus importante pour les réponses auto-administrées que pour les réponses par téléphone : l'étendue du vocabulaire est deux fois plus importante en auto-administré ; on a 2 370 mots distincts alors que, par téléphone, on n'en a que 874.

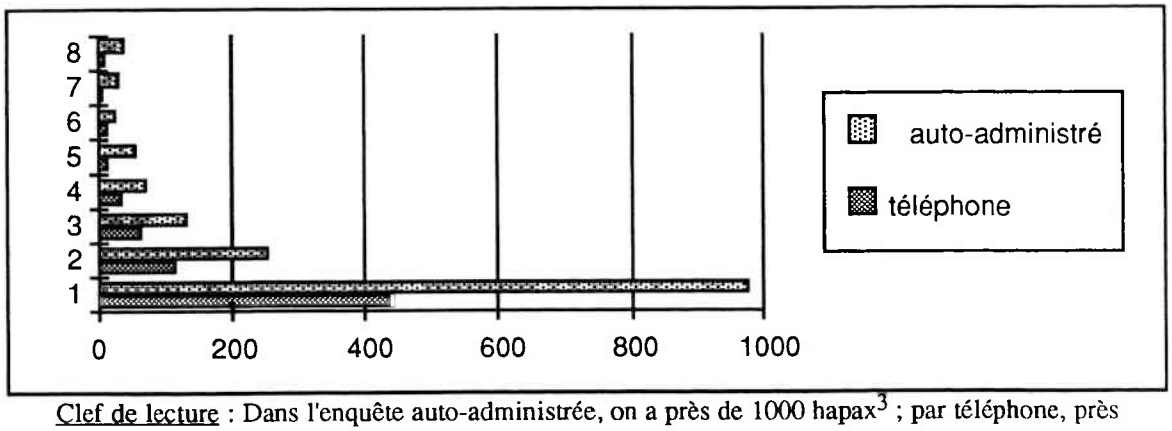
D'où vient cet écart ? L'allure des courbes de Zipf permet de localiser les fréquences d'occurrences pour lesquelles les écarts sont les plus importants. Grâce à cette courbe, on peut voir comment se répartissent les mots en fonction de leur fréquence d'utilisation. Zipf (1935) a montré que sur n'importe quel corpus de texte on retrouvait des courbes d'allures similaires. Si l'on classe les mots par fréquence décroissante, très peu de mots apparaissent avec une fréquence élevée, et de plus en plus de mots apparaissent avec une très faible fréquence. Sur nos deux corpus, on retrouve bien la même allure générale mais il existe des différences sensibles entre les deux courbes de répartition.

Graphique 1 : Courbes de Zipf sur les deux corpus de réponses (auto-administré, panel)



Dans une large mesure, l'écart est dû aux mots ayant un très faible nombre d'occurrences. Pour les fréquences de mots élevées, on a à peu près autant de mots en auto-administré que par téléphone. En revanche, pour les faibles fréquences, le nombre de mots est deux fois plus élevé en auto-administré comme le montre le graphique ci-après.

Graphique 2 : Nombre de mots de basse fréquence (détail de la courbe de Zipf)



Clef de lecture : Dans l'enquête auto-administrée, on a près de 1000 hapax[3] ; par téléphone, près de 500 hapax.

[3] Hapax : mot n'apparaissant qu'une fois dans le corpus.

En auto-administré, les réponses sont plus longues et utilisent un vocabulaire beaucoup plus varié. C'est la première différence majeure entre les deux modes de questionnement. Par écrit, les personnes interrogées ont le temps de construire leur représentation mentale de l'objet (une représentation plutôt conceptuelle ou plutôt descriptive) et de choisir les mots les mieux adaptés à l'énonciation de cette représentation. Il en découle logiquement des réponses plus longues et plus riches.
Notons au passage que le temps est sans doute un facteur déterminant pour l'expression de la part la plus idiosyncrasique de la représentation mentale. Placé dans une situation d'urgence par l'appel téléphonique, l'enquêté aura tendance à donner une réponse courte qui aura toutes les chances de reproduire des stéréotypes (culturels, publicitaires ...). Nous avions déjà observé ce phénomène dans l'étude sur la qualité. Dans les réponses aux questions ouvertes sur la notion de qualité à propos de certains produits ou services, on retrouvait des éléments du discours publicitaire (Beaudouin, Hébel et Lahlou, 1992).

Il semble que l'enquêté interrogé par téléphone s'investisse peu dans sa réponse. Interrompu dans ses activités quotidiennes par l'appel téléphonique, il n'a probablement qu'un souci, celui de se débarrasser au plus vite de l'enquêteur. L'enquêteur, de son côté, essaie de terminer l'enquête au plus vite pour améliorer son rendement.
Par écrit, on obtient les réponses de personnes suffisamment motivées pour prendre le temps de réfléchir, d'écrire et de renvoyer le questionnaire. L'enquêté est dans une situation de liberté maximale face à sa feuille de papier, tandis qu'en liaison téléphonique avec l'enquêteur, il sent beaucoup plus peser sur lui la contrainte, la nécessité de répondre.

Rappelons enfin que les panélistes de la SÉCODIP sont "rémunérés" en nature (bons d'achats), ce qui peut les motiver à bien répondre. Interrogés par téléphone, ils auraient peut être donné des réponses plus élaborées.

1.2.2. Réduire la taille du vocabulaire : atténuation des différences

En analyse des données textuelles, on travaille sur des matrices qui contiennent :

- en ligne : les réponses ou des fragments de réponses ;
- en colonne : des occurrences de mots.

Pour que l'on puisse comparer les profils lexicaux des réponses avec une bonne robustesse statistique, on diminue le nombre de mots entrant dans l'analyse. La réduction de la taille du vocabulaire consiste à supprimer les mots apparaissant moins de quatre fois et à lemmatiser. Ceci permet d'extraire de cette matrice en général très creuse, une matrice plus dense.

Ensuite, puisque l'analyse lexicale est indifférente à la syntaxe, on décide de traiter en variables illustratives les mots-outils (ou mots grammaticaux). Cette dernière étape permet de réduire encore la taille du vocabulaire analysé. Mais cette réduction de taille a tendance à modifier la nature des corpus analysés : elle accentue les phénomènes de masse et efface les singularités. Cette étape a notamment pour effet de réduire les écarts entre les deux types de corpus comme le montre le tableau suivant.

Tableau 1 : Comparaison de la richesse lexicale sur les deux corpus et effet de la réduction

	Téléphone		Auto-administré	
	Avant réduction	Après réduction et reconnaissance des mots-outils	Avant réduction	Après réduction et reconnaissance des mots-outils
Nombre de MOTS DISTINCTS	874	232	2 370	453
dont nombre d'hapax	440	0	979	0
Nombre de MOTS ANALYSÉS (mots pleins)		165		317
Nombre de MOTS ILLUSTRATIFS (mots-outils)		67		136
Nombre moyen de MOTS PAR RÉPONSE	15,4	13,4	18,7	13,4

Après traitement, on constate d'une part que la taille du vocabulaire a été réduite par quatre pour les réponses téléphoniques, par cinq pour les réponses auto-administrées. Les trois quarts des mots employés n'interviennent plus dans l'analyse. Plus la taille du vocabulaire initiale est grande, plus les effets de la réduction sont importants. On a vu sur les deux corpus que l'écart maximal au niveau de la répartition du vocabulaire se situait dans les faibles fréquences. En éliminant les mots peu employés, on a tendance à réduire l'écart absolu entre les deux corpus.

D'autre part, en éliminant les mots-outils, on travaille sur moins d'un cinquième du vocabulaire initial. L'élimination des mots-outils a tendance à être un peu plus défavorable aux textes écrits. En effet, la proportion des mots-outils est plus importante en auto-administré : 43 % contre 41 %. Ceci est assez logique : dans la rédaction écrite le recours à l'appareillage syntaxique est beaucoup plus naturel que dans le langage parlé. Dans le discours oral, la coordination ou la juxtaposition se substituent facilement à la subordination. En fin de course, pour le questionnaire auto-administré, le vocabulaire analysé ne représente plus que 13 % du vocabulaire initial, alors qu'il représente 19 % du vocabulaire de départ pour l'enquête téléphonique.

La réduction a une forte incidence sur le nombre de mots analysés dans les réponses. Après suppression des mots apparaissant moins de quatre fois, l'écart entre le nombre moyen de mots sur les deux échantillons a disparu. Dans les deux cas, on a 13,4 mots en moyenne. La réduction a eu un effet beaucoup plus fort sur les réponses écrites puisque l'on a une diminution de cinq mots. La réduction a pour effet d'atténuer fortement les singularités pour mieux mettre en évidence les effets de masse. Pour paraphraser Touraine[4], "la statistique est l'ennemie du moi".

En résumé, on a vu que les réponses écrites sont beaucoup plus longues, puisent dans un vocabulaire plus ample et ont davantage recours aux outils syntaxiques que les réponses orales. Ceci est en partie dû au respect des règles d'expression du discours (respect de la syntaxe) qui pèse plus dans la rédaction que dans l'expression orale. La réduction de la taille du vocabulaire indispensable à l'analyse lexicale a, quant à elle, tendance à réduire les différences entre l'écrit et l'oral puisqu'elle élimine les mots de faibles fréquences et les outils syntaxiques qui sont plutôt caractéristiques du discours écrit.

[4] "La sociologie est l'ennemie du moi" in (Touraine, 1974).

1.3. Les différences de contenu : surface téléphonique, profondeur de l'écrit

Les différences entre les réponses auto-administrées et téléphoniques ne tiennent pas seulement à des variations de richesse lexicale et syntaxique, elles touchent au sens. Ce sont ces variations en terme de contenu sémantique que nous analysons à présent.

1.3.1. Les mots les plus fréquents

Le simple examen des mots les plus fréquents, - les mots cités par plus d'une personne sur dix - donne un aperçu des différences de contenu dans les réponses.

1.3.1.1. Les points communs

Sur les deux corpus, ce sont les produits composant le petit déjeuner qui sont les plus souvent cités (description pragmatique, en termes d'objets-du-monde).
Ce sont les mêmes noms de produits qui reviennent le plus souvent, même si l'ordre n'est pas tout à fait le même. L'ensemble café, pain, beurre, confiture vient en tête sur les deux sous-échantillons, comme s'il s'agissait du prototype de petit déjeuner le plus répandu. Notons que par téléphone, les mots les plus fréquents ont tendance à être encore plus souvent répétés que par écrit. *Beurre* revient deux fois plus souvent par téléphone que par écrit, et la fréquence des termes *café, confiture, pain* est elle aussi plus élevée. (Notons aussi que *croissant* est cité par un consommateur sur cinq.)

Pour le jus de fruit, on note un écart assez important entre les deux corpus. En auto-administré, 40 % des personnes citent le *jus de fruit* (et en particulier le jus d'orange), alors que par téléphone seuls 20 % le citent. Il semblerait que le jus de fruit soit un composant moins fondamental du petit déjeuner. S'il n'est pas véritablement ancré dans les habitudes de consommation au petit déjeuner - il ne figure que dans 5 % des petits déjeuners français d'après une étude de Taylor Nelson - le jus de fruit est une des composantes privilégiées de l'image publicitaire du petit déjeuner.

La représentation du petit déjeuner idéal serait davantage influencée par l'image publicitaire quand l'enquêté a le temps de répondre, tandis que les réponses téléphoniques, moins idiosyncrasiques, se rapprochent du stéréotype "boisson chaude plus tartine". Nous n'avons pour le moment pas les moyens de vérifier plus en détail cette hypothèse.

Tableau n° 2 : Les mots les plus fréquents (en auto-administré et par téléphone)

Auto-administré	
336	***journée***
281	café
251	pain
237	**bon**
202	beurre
195	confiture
186	***famille***
174	lait
161	croissante
156	**bien**
156	***temps***
155	***moment***
115	***faire.***
107	jus de fruit
102	grillé
100	chocolat
99	frais
97	***repas***
94	jus d'orange
93	***commenc+***
93	***détend+***
90	***vacances***
89	***calme***
89	fruit
86	***table***

Téléphone	
413	beurre
361	café
344	confiture
343	pain
278	jus d'orange
255	tartine
197	croissant
191	lait
171	café au lait
166	céréale
139	oeuf
135	fruit
135	jus de fruit
134	chocolat
118	grillé
106	thé
83	**bon**

Légende
Les termes en "police standard" correspondent aux produits.
Les termes en "gras" aux caractérisants hédoniques.
Les termes en "italique" aux termes caractérisant le contexte, l'ambiance, le lieu, ...

Nota : La taille des échantillons étant presque identique, nous n'avons pas pondéré les fréquences par la taille de l'échantillon.

1.3.1.2. Les différences : structure syntaxique et contenu sémantique

• La liste des mots-outils les plus fréquents confirme bien ce que nous disions précédemment. Ils sont beaucoup plus nombreux et fréquents en auto-administré.

Tableau n° 3 : Les mots-outils les plus fréquents (en auto-administré et par téléphone)

Auto-administré	
507	et
331	ou
314	en
289	pour
283	avec
148	dans
143	je
117	c'est
116	se
113	que
106	me
106	on
104	qui
98	ne

Téléphone	
184	avec
178	et
99	ou
86	en

Les conjonctions de coordination ont tendance à disparaître dans le discours oral retranscrit (c'est sans doute davantage le fait de la retranscription que du discours oral lui-même).

Le *je* n'apparaît pratiquement pas dans les réponses par téléphone, alors qu'en auto-administré, 14 % des enquêtés l'emploient. Est-ce lié au caractère oral ou à la retranscription ? Par téléphone, la présence de l'énonciateur est tellement évidente qu'elle n'a pas besoin d'être signalée. En auto-administré, le respect des règles syntaxiques pousse à employer le *je* plus fréquemment.
Enfin, les conjonctions de subordination pratiquement absentes de l'enquête téléphonique sont très fréquentes dans les réponses auto-administrées.

Tout se passe comme si, en passant de l'écrit à l'oral, on passait du syntagme nominal avec subordonnées au syntagme nominal simple. La structure phrastique avec subordonnées cède la place à une juxtaposition de syntagmes nominaux.

Réponses écrites :

* Aux périodes de vacances car c'est là qu'on l'apprécie le plus. On a le temps. Idéal veut dire avec le temps le soleil dehors un pain très frais et chaud, et encore mieux à deux.
* À quelque chose de bien préparé, d'attendu, c'est un moment privilégié qui prépare la journée, que l'on ne devrait jamais manquer car c'est très important et cela évite les coups de pompe de la matinée

Réponses orales :

* Jus d'orange, corn flakes, beurre, pain, laitage, fruit, c'est tout.
* Lait, céréales, fruit, c'est tout.
* Je n'en prends pas. Sinon je suis pour les petits déjeuners anglais ou américains, du fromage, oeuf, des jus, du bacon, des saucisses, du jambon blanc ou fumé.

* Une orange pressée ou un pamplemousse pour la vitamine C, du lait avec du cacao, une tartine de beurre ou de confiture et un yaourt.

L'analyse comparée de l'effet de l'écholalie[5] montre aussi les différences dans le traitement de la syntaxe entre l'oral et l'écrit. En réponse à une question ouverte, on a souvent remarqué que les mots les plus fréquents dans les réponses correspondent aux termes employés dans la question. Nous avons donc pris l'habitude d'exclure de l'analyse les termes contenus dans la question en partant de l'idée qu'ils n'apportaient aucune information supplémentaire.

Sur nos deux corpus, on remarque ce phénomène d'écholalie, mais il est beaucoup moins important pour l'enquête téléphonique. On a cinq fois moins d'occurrences des termes *petit déjeuner* et *idéal* que dans les réponses écrites.

Phénomène d'écholalie

	Auto-administré	Téléphone
Idéal	94	15
Petit déjeuner	263	54
Petit	66	36
Déjeuner	65	29

Les différences peuvent s'expliquer de deux manières :

- il est possible que l'enquêteur téléphonique, en dépit des consignes, ne note pas ces parties de la réponse qui, à ses yeux, ne contiennent pas d'information. Il aurait tendance à faire un prédécoupage sur les réponses.

- le discours écrit conduit à répondre à la question de manière complète, c'est-à-dire en rappelant le thème avant d'entrer dans l'opération de prédication. Et nous retrouvons là un effet de l'apprentissage scolaire. L'enquête par écrit est assimilée dans l'esprit des enquêtés à une épreuve scolaire, d'autant plus qu'il s'agit de personnes régulièrement interrogées. Le panéliste est un peu comme l'élève face au maître avec lequel il reste une année entière, avec le désir de "bien faire".

Pour les panélistes, on retrouve bien ce phénomène d'écholalie. Exemple :

* Le petit déjeuner idéal, c'est être nombreux autour d'une table bien remplie, pain, brioche, croissant, beurre, miel, confiture, thé ou café, chocolat, yaourt, orange.

[5] L'écholalie est un phénomène de répétition qui se propage de la question à la réponse.

De plus, par téléphone, quand l'expression 'petit déjeuner' est employée, ce n'est pas dans un souci de thématisation (rappeler l'élément à définir), mais simplement comme support d'une caractérisation :

> * Un petit déjeuner anglais, oeufs, bacon, tomate, toast grillé, confiture d'orange ou citron / dans un jardin au soleil, sous un parasol, bien servi.
> * Petit déjeuner traditionnel : pain beurre, croissant, chocolat ou café au lait, assez copieux.

La répétition des termes de la question est beaucoup plus importante dans le discours écrit, c'est un embrayeur de production de réponse qui complexifie la structure syntaxique de la phrase. Elle est pratiquement inexistante dans les réponses orales retranscrites.

• Outre les noms de produits, sur le corpus des réponses auto-administrées, on trouve, parmi les mots ayant une fréquence élevée, des termes qui ont trait au cadre du petit déjeuner. Il semble que la prise en compte du contexte social, géographique, de l'ambiance et de la place du petit déjeuner par rapport au déroulement de la journée n'apparaisse de façon significative que dans les réponses écrites.

L'apparition de termes autres que des noms de produits est liée à un changement de focalisation ou de point de vue. Dans la focalisation locale, le petit déjeuner est un ensemble d'ingrédients que l'on met sur la table, que l'on prépare et/ou que l'on consomme. Le petit déjeuner est considéré indépendamment de sa fonction énergétique, sociale, morale... Il est possible que le contexte d'interrogation ait ici joué. En l'absence d'une direction précise orientant les associations, l'enquêté risque de rester sur un noyau minimal, et donc stéréotypé. En auto-administré, la présence d'autres questions sur le cadre à côté de la question étudiée a dû influencer également les enquêtés. Par ailleurs, le cadre écrit suscite des réponses longues (syndrome du bon élève) et amène donc l'enquêté à rechercher un enrichissement du noyau de réponse minimal, enrichissement qui a de fortes chances d'être non pragmatique (puisque le stéréotype est ici pragmatique) et idiosyncrasique, par construction.

Les réponses en termes de noms de produits correspondent à ce que, dans l'étude sur la qualité (Beaudouin, Hébel et Lahlou, 1992), on avait appelé "réponses prototypiques". En revanche, les termes liés à l'ambiance, au cadre, ... traduisent une vision plus globale du petit déjeuner. Moment particulier de la journée et éventuellement de la vie sociale, il joue aussi un rôle dans le déroulement de la journée. L'esprit ne se contente plus d'énumérer les ingrédients situés sur la table mais de voir aussi ce qu'il y a autour dans l'espace (famille, musique, radio...) et dans le temps (point de départ de la journée...). La description est alors plus multimodale.

Les réponses plus globales correspondent à une tentative de conceptualisation.
En résumé, les réponses qui se présentent comme une énumération de noms de produits correspondent à une approche locale et prototypique, tandis que les autres correspondent à une approche globale et conceptuelle. On reviendra plus loin sur cette opposition entre les réponses prototypiques et les réponses conceptuelles.

Finalement, il est clair que le questionnaire auto-administré conduit à des réponses plus conceptuelles et plus générales de ce qu'est le petit déjeuner idéal, ce qui n'exclut pas qu'au cours de la réponse, les personnes citent aussi des noms de produits à titre d'exemple.

Pourquoi le conceptuel et le général sont-ils moins fréquents dans l'enquête téléphonique ? Le passage du stade descriptif au stade conceptuel nécessite un effort de pensée, il faut se détourner du sensible pour atteindre l'intelligible, comme le faisait entendre Socrate. Or la situation d'urgence liée à l'appel téléphonique ne s'y prête guère.

Nous avions déjà constaté que le vocabulaire de l'écrit est plus riche que celui de l'oral. Pour la syntaxe, la même observation s'impose : les structures syntaxiques de l'écrit ont tendance à être beaucoup plus complexes et variées. Enfin, les représentations mentales associées au petit déjeuner sont plus variées, en ce sens que le cadre de référence est souvent beaucoup plus large. En plus des énumérations de noms de produits, on a des références au contexte du petit déjeuner et des tentatives de réponses synthétiques ou conceptuelles.

Cette complexité n'a pas forcément une valeur supérieure. La simplicité a une efficacité pragmatique et communicative évidente. On peut penser que la présence immédiate de l'enquêteur joue un rôle, l'enquêté cherchant à lui faire comprendre le plus rapidement possible son point de vue. La communication sur les objets (produits) est plus immédiate que la communication sur les concepts.

1.3.2. Typologie des réponses

On se propose pour vérifier les hypothèses nées du simple examen des mots les plus fréquemment employés, d'analyser simultanément les deux corpus de réponses à l'aide d'une classification descendante hiérarchique : l'ensemble des énoncés est d'abord segmenté en deux classes maximisant les différences de répartition du vocabulaire. Autrement dit, chaque classe est aussi homogène que possible en terme de vocabulaire employé alors que les distinctions

entre les deux classes sont aussi marquées que possible. La plus grande des classes restantes est à son tour segmentée, etc. ...

On obtient ainsi une typologie des discours propres au petit déjeuner dont on présentera ici les grandes lignes (une analyse plus détaillée figurera dans un prochain rapport). Ici, on s'attachera surtout à l'étude de l'incidence du mode de questionnement. Si au premier nœud de découpage les réponses auto-administrées se distinguent des réponses par téléphone, on pourra conclure que le mode de questionnement a une forte incidence sur les types de réponses. Rappelons qu'il y a un biais dans l'échantillon dû au fait que les panélistes sont exclusivement des femmes. On cherchera donc à départager l'incidence du mode de questionnement et du sexe de l'enquêté sur les représentations.

Déjeuner vient du verbe déjeuner (rompre le jeûne) substantivé. Il désigne l'action de se nourrir, de prendre un repas et, par métonymie, les mets qui le constituent et le moment du déjeuner.

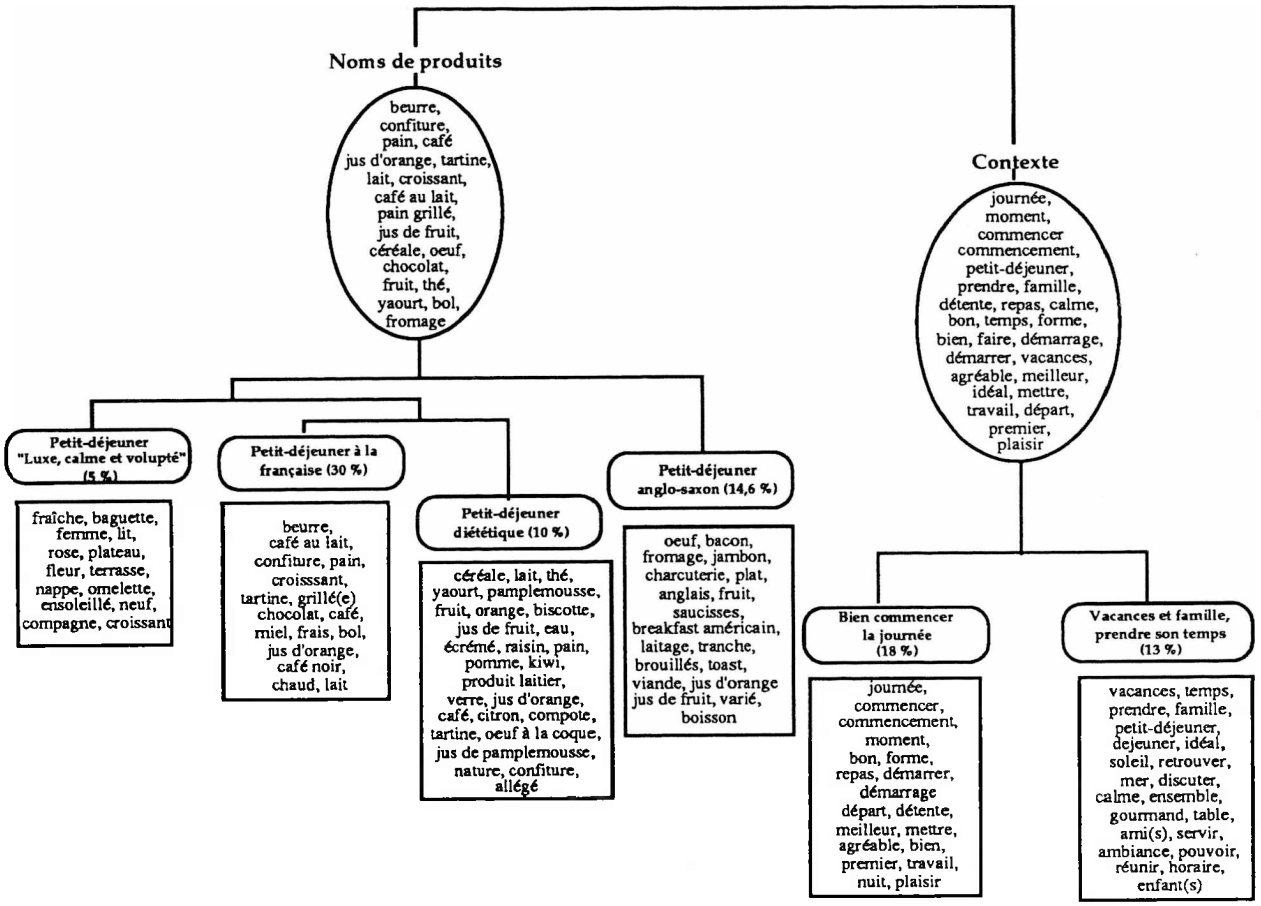
L'analyse lexicale met en évidence deux manières très distinctes de répondre à la question qui correspondent à deux manières de l'envisager :

- en énumérant les noms de produits, c'est la fonction alimentaire, biologique du petit déjeuner qui prime. Mais ces énumérations nous informent aussi sur les habitudes culturelles du petit déjeuner ;

- en évoquant son contexte spatial, temporel et social, le petit déjeuner est alors envisagé comme un moment particulier de la journée associé à certains rites sociaux du réveil (convivialité, informations...).

Les évocations nous fournissent six classes qui se répartissent en deux grandes surclasses : les noms de produits et le contexte non alimentaire.

Graphique 3 : Les évocations associées au petit déjeuner idéal (téléphone plus auto-administré)



1.3.2.1. Les constituants du petit déjeuner

Parmi les réponses en terme de noms de produits, on distingue trois grandes classes qui correspondent à trois types de petits déjeuners. Une quatrième classe a une place intermédiaire entre le pôle noms de produits et le pôle contexte.

Les trois types de petit déjeuner sont :

a) Le petit déjeuner continental ou "à la française" qui correspond à 30 % des réponses : café (noir ou lait), pain et produits panifiés, beurre, confiture.

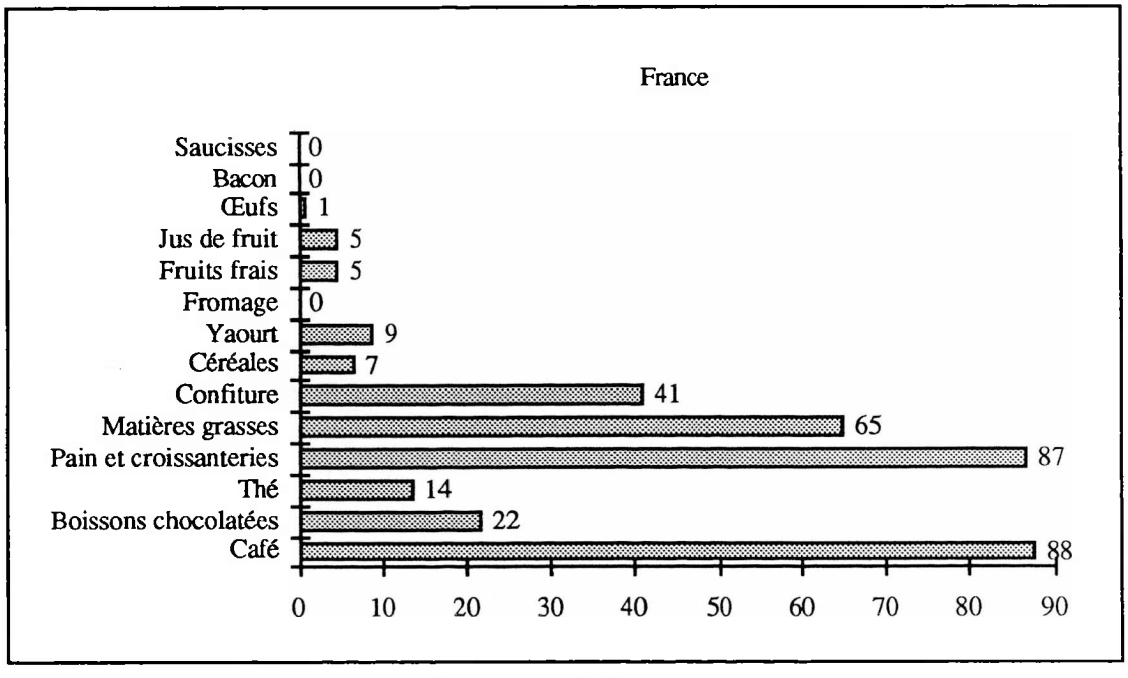
Réponses caractéristiques :

* Une tasse de café avec un croissant ou du pain grillé, tartine de confiture
* Café au lait, tartines de pain frais avec beurre et confiture de fraises

* Café, tartines grillées, pain, confiture, beurre.
* Chocolat chaud / croissants / tartines / beurre / confiture
* Un bol de café / avant tartines de beurre et de confiture / maintenant juste un café
* Café au lait, croissants, pain grillé, beurre, confiture.

Ces descriptions correspondent assez bien aux petits déjeuners typiques en France. Comme le montre le graphique ci-après, le café est en effet la boisson chaude privilégiée du matin, bien plus répandue que le thé (contrairement aux Anglo-saxons) ou que le chocolat chaud. Le café est en effet consommé par 88 % des Français. Le pain-beurre-confiture vient aussi en tête.

Graphique 4 : Les constituants du petit déjeuner français - panel Taylor Nelson



En % des petits déjeuners

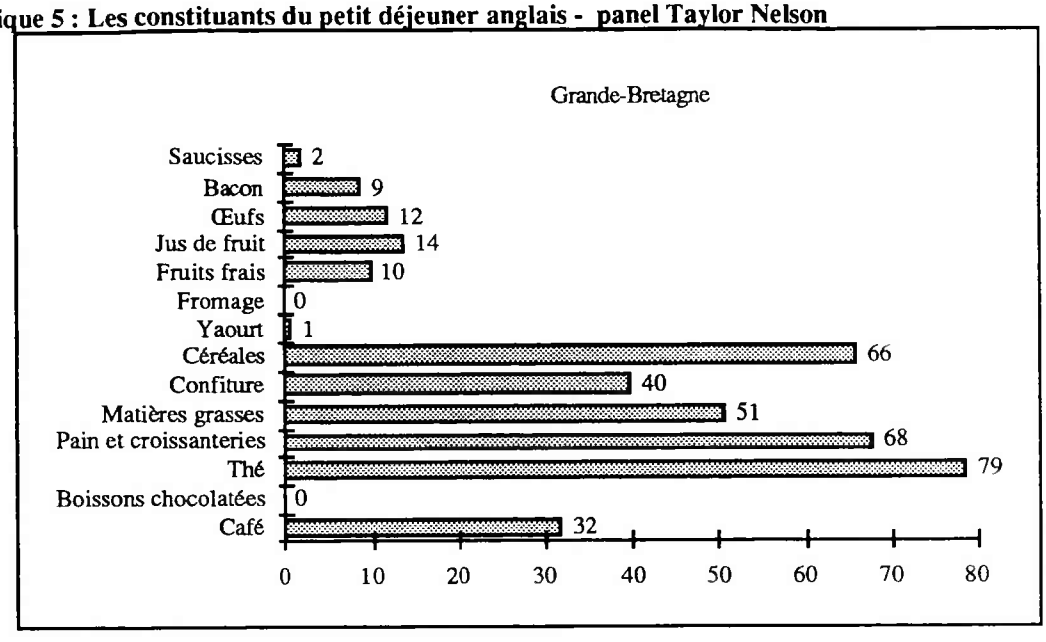
b) Le petit déjeuner diététique est un idéal pour 10 % des personnes interrogées. Il est constitué à base de fruits (jus de fruits, fruits frais, compote), produits laitiers (lait, yaourt) et céréales. On remarquera l'absence significative de produits de charcuterie. La boisson chaude qui accompagne le plus fréquemment ce type de petit déjeuner est le thé et non le café. Ce type de petit déjeuner est conforme aux recommandations des diététiciens et des magazines féminins sensibilisés aux problèmes de la forme dans les années 80.

Réponses caractéristiques :

* Thé citron, céréales, lait, fruit, yaourt, dattes, mandarine, figues.
* Un jus de pamplemousse, un kiwi, des céréales, des toasts et un produit laitier.
* Lait avec des céréales et un jus de fruit, biscotte avec confiture.
* Céréales / grand bol de thé au lait / un jus de fruits / (pamplemousse / abricot / orange) une tartine / un oeuf à la coque.
* Gâteau de riz au caramel, une tasse de thé sans sucre, cinq cuillerées à soupe de céréales soufflées avec un verre de lait écrémé plus une orange.
* Céréales, verre de jus de fruit.
* Jus de fruit, céréales, lait.
* Un demi pamplemousse, deux tartines de confiture, un yaourt nature maigre, thé.

c) Le petit déjeuner à l'anglo-saxonne avec des œufs, de la charcuterie, du fromage et éventuellement des fruits. Il est désigné par son nom générique par la plupart des personnes interrogées : *breakfast, petit déjeuner anglais, petit déjeuner américain* ... Notons que pour les personnes interrogées, le fromage est un des constituants du petit déjeuner anglais ou américain alors que, dans la pratique, il n'apparaît pas dans le breakfast comme le montre le graphique ci-après :

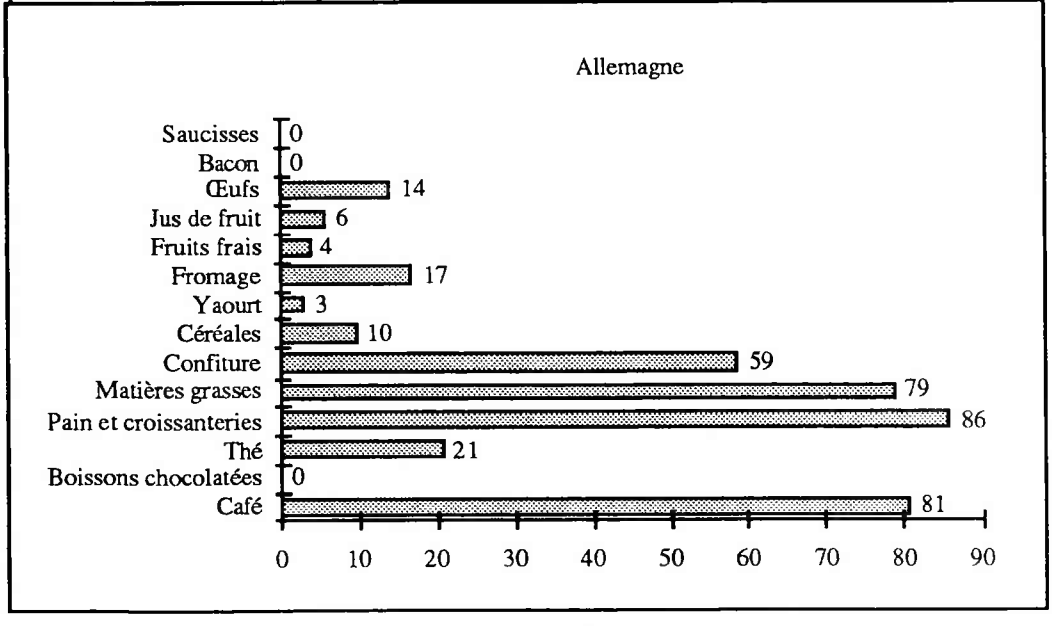
Graphique 5 : Les constituants du petit déjeuner anglais - panel Taylor Nelson



En % des petits déjeuners

Il s'agit donc bien d'une représentation que les Français se font des repas de leurs voisins qui n'a qu'un ancrage partiel dans la réalité des pratiques. Il ressemble davantage à une combinaison des petits déjeuners anglo-saxon et germanique à base de produits panifiés, de charcuterie et de fromage.

Graphique 6 : Les constituants du petit déjeuner allemand - panel Taylor Nelson



En % des petits déjeuners

Réponses caractéristiques :

* Un jus d'orange, des oeufs brouillés avec bacon ou jambon, fromage, toasts, confiture, thé
* Oeuf, bacon, toast, vitamine, jus de fruit / costaud, style anglais.
* Breakfast anglais, jus d'orange, toast, gâteaux, fromage, charcuterie.
* Je n'en prends pas / sinon je suis pour les petits déjeuners anglais ou américains, du fromage, oeuf, des jus, du bacon, des saucisses, du jambon blanc ou fumé.
* A l'américaine, qu'on puisse avoir le choix entre de nombreux produits / oeufs, viande, yaourts, fruit, fromage, jambon.
* Oeuf au bacon, porridge, jus d'orange, tisane, fruit, potage.
* Jus d'orange, toasts grillés, différentes confitures, jambon, oeufs sur le plat.

Ces trois grands types correspondent à ceux trouvés par une méthode complètement différente dans l'enquête sur les comportements alimentaires du CREDOC (Lahlou, 1989 ; Yvon, 1989). Nous y reviendrons en détail dans la partie suivante.

Dans les descriptions du petit déjeuner idéal que nous venons d'examiner, les réponses peuvent facilement être rattachées à des prototypes bien connus du petit déjeuner. Si l'idéal ne correspond pas aux pratiques effectives, on peut cependant noter qu'il ne se distingue pratiquement pas de modèles de petits déjeuners culturellement attestés. L'idéal n'est qu'un prolongement de la réalité, sans rupture.

d) Le petit déjeuner "Luxe, calme et volupté"

Du côté des noms de produits, on a encore une classe de faible effectif intéressante de par son contenu. Elle occupe une position intermédiaire dans la mesure où il s'agit d'une énumération de produits mais contient également une référence au contexte. Le principe de plaisir est au cœur de ces réponses signifié par de nombreux adjectifs hédoniques (bon, beau, joli...) :

- le plaisir des yeux : beau panorama, vue dégagée ("une terrasse ensoleillée", "au bord d'une piscine", "avec vue superbe") ; belle table ("une jolie nappe", "sur un plateau magnifiquement décoré").
- le plaisir gustatif : produits frais ("baguette bien fraîche") ;
- le plaisir d'être servi ("hôtel", "servi par une belle femme") ;
- le plaisir d'être en bonne compagnie ("avec des jolies femmes", "avec mon mari", "une bonne compagnie, ma femme").

Il s'agit en quelque sorte de petits déjeuners "enrichis" de produits consommables mais non comestibles.

Réponses caractéristiques :

* Fruits de Madagascar, thé / sur une plage ensoleillée, avec des jolies femmes.
* Café au lait, croissants frais ou baguette fraîche (salée), tout cela pris sur une terrasse avec une belle vue.
* Café, croissant, brioche / dans un hôtel ça me plairait, j'apprécie en plus avec mon mari.
* Café à la montagne / en terrasse avec un beau soleil avec ma femme / un bon jus de fruit.
* Jus d'orange avec croissants / dans un hôtel, on m'amène tout sur un plateau, au lit.
* De préférence dehors au jardin ou sur une terrasse avec une jolie nappe, des grands bols, pain grillé, baguette fraîche, croissants avec thé, café, lait et chocolat pour les enfants, jus de fruit.
* Qu'il me soit servi avec pain frais, croissants, brioches, lait, café. Ce, dans un bel endroit, avec vue superbe : montagne ou mer peu m'importe, mais beau panorama.
* Dans un hôtel, pas au lit, sur une table fleurie avec des couverts en argent, des tasses en porcelaine style anglais / petit déjeuner continental.

1.3.2.2. Le contexte du petit déjeuner

Cet ensemble de réponses correspond à un tiers de l'échantillon global et se scinde en deux classes, l'une centrée sur la fonction du petit déjeuner (réservoir d'énergie pour la journée), l'autre sur la manière dont il est pris (compagnie, lieu, durée...). Le point commun entre toutes ces réponses est que le petit déjeuner idéal évoque un moment de la journée et non pas seulement l'action de se nourrir. Les produits consommés sont une des composantes d'un tout plus vaste qui est soit la journée, soit une partie de la journée.

a) Le point de départ de la journée

Le petit déjeuner idéal évoque à un moment agréable de la journée qui, de par sa situation liminaire, conditionne le bon déroulement de celle-ci.

Réponses caractéristiques :

* Un agréable moment à vivre sans être trop pressé, dans la détente qui permet de bien démarrer la journée, c'est un peu comme un lever de rideau sur une scène de théâtre : on est d'emblée séduit par le décor, et donc bien disposé à accueillir la suite, ou bien mal à l'aise et donc sur la défensive.
* A bien démarrer la journée, se sentir en forme, et faire le plein d'énergie, éviter le coup de pompe en fin de matinée.
* Le petit déjeuner est la chose qui peut nous mettre de bonne humeur pour la journée aussi faut-il qu'il se passe dans de bonnes conditions, il ne me fait penser à rien en général mais c'est quand même un moment très important.
* A un moment de plaisir, pour démarrer une journée dans les meilleures conditions. C'est le moment le plus agréable de la journée.

* Moment agréable, première occupation de la journée, c'est la mise en route lente des activités journalières.
* Vacances / agréable / plaisir une bonne fin de nuit / bar / démarrage de la journée / une bonne chose de faite.
* A quelque chose de bien préparé, d'attendu, c'est un moment privilégié qui prépare la journée.
* Que l'on ne devrait jamais manquer car c'est très important et cela évite les coups de pompe de la matinée.

b) Une occasion de convivialité familiale

Dans cette classe, le petit déjeuner idéal ne fait pas penser aux mets qui le composent mais au moment privilégié qu'il peut signifier. C'est un prétexte pour se retrouver entre soi, prendre le temps de discuter, profiter des autres. Occasion privilégiée de la sociabilité familiale, de l'échange, sa durée revêt une grande importance pour les enquêtés de cette classe. Le petit déjeuner idéal fait penser aux vacances -moment où peuvent être réunies les conditions idéales pour l'échange familial-.

Réponses caractéristiques :

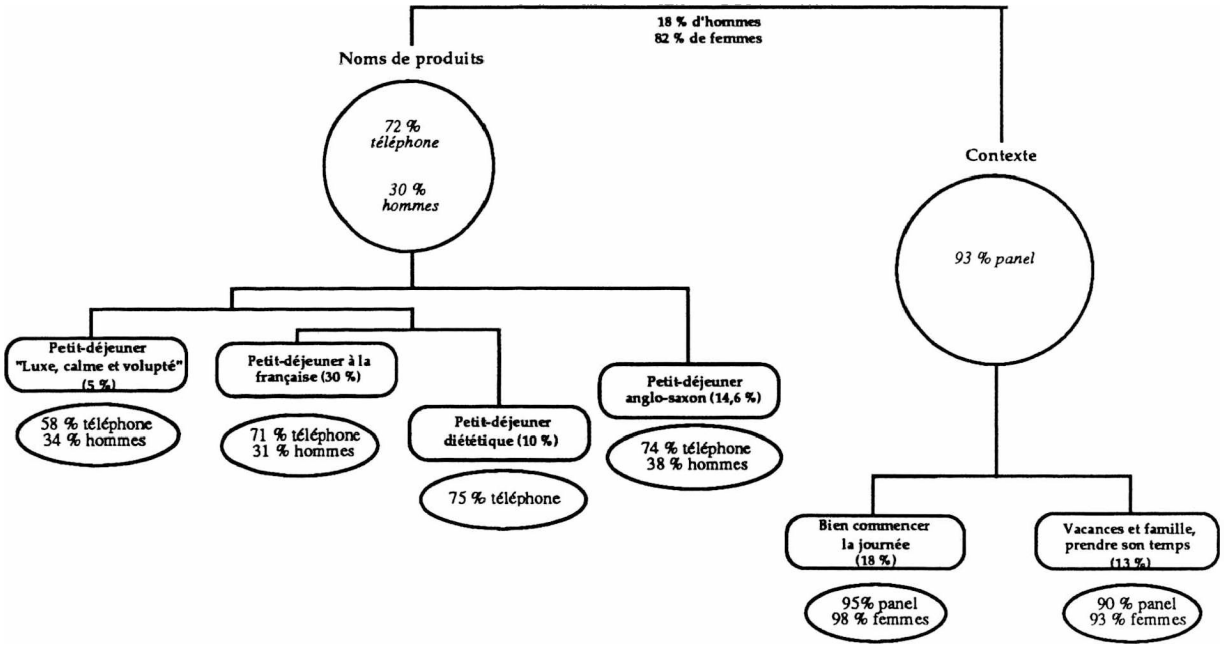
* Aux vacances, c'est un des rares moments où l'on prend son temps pour le petit déjeuner, détendu, au calme, instants délicieux, sur une terrasse, dans le jardin, où l'on savoure ce que l'on a envie de manger, où l'on prend le temps d'écouter, à partager avec ceux que l'on aime.
* Un moment privilégié où toute la famille est réunie, c'est-à-dire le plus souvent le dimanche matin en prenant son temps et pouvoir enfin discuter ensemble de notre semaine calmement après ces journées assez stressantes.
* Un petit déjeuner idéal est un petit déjeuner en famille : parents, enfants, grands-parents, c'est ainsi pendant les vacances, on déjeune, on parle du film ou de l'émission de la veille, ou on se raconte nos rêves, on prend son temps, on feuillette le journal et on commente.
* Cela me fait penser aux vacances, car c'est la seule période où on se retrouve tous les matins en famille. On prend le temps de faire un bon petit déjeuner copieux avec tout ce dont on a envie de manger.
* Le petit déjeuner idéal, c'est lorsque l'on ne travaille pas, que l'on n'est pas pressé, prendre le temps de beurrer les brioches, les étendre tranquillement de confiture, l'odeur du café chaud.

1.3.3. Incidence du mode de questionnement

Nous venons de caractériser les grandes dimensions des évocations associées au petit déjeuner idéal. Il nous reste à voir dans quelle mesure le type de questionnement a eu une influence sur les réponses.

Dès la première segmentation, les réponses du panel se distinguent très nettement des réponses téléphoniques. Une première classe qui correspond à un tiers du corpus est constituée à 93 % par des réponses du panel : il s'agit des réponses qui ont trait au cadre et à la convivialité (contexte). Les trois quarts des réponses de la seconde classe (64,5 % des réponses) proviennent de l'enquête téléphonique et correspondent à des énumérations de noms de produits constituant le petit déjeuner.

Graphique 7 : Typologie des représentations et mode de questionnement



Clef de lecture :
La classe "petit déjeuner à la française" représente 30 % des réponses (auto-administrées plus téléphoniques). Parmi ces réponses, 71 % proviennent de l'enquête téléphonique, 29 % des questionnaires auto-administrés.
31 % des réponses proviennent d'hommes, alors que les hommes représentent 18 % de l'échantillon global.

Il apparaît donc que les réponses qui prennent en compte le contexte et l'entourage sont pratiquement absentes des réponses téléphoniques (5 % seulement). En revanche, sur le panel, on obtient à la fois des réponses en terme de noms de produits (38 %) et des réponses qui évoquent le contexte (62 %). La "représentation sociale" du petit déjeuner qui transparaît dans le panel est donc beaucoup plus variée que celle de l'enquête téléphonique.

Mais ces différences sont-elles dues au fait qu'il n'y a pas d'hommes dans l'échantillon auto-administré ?

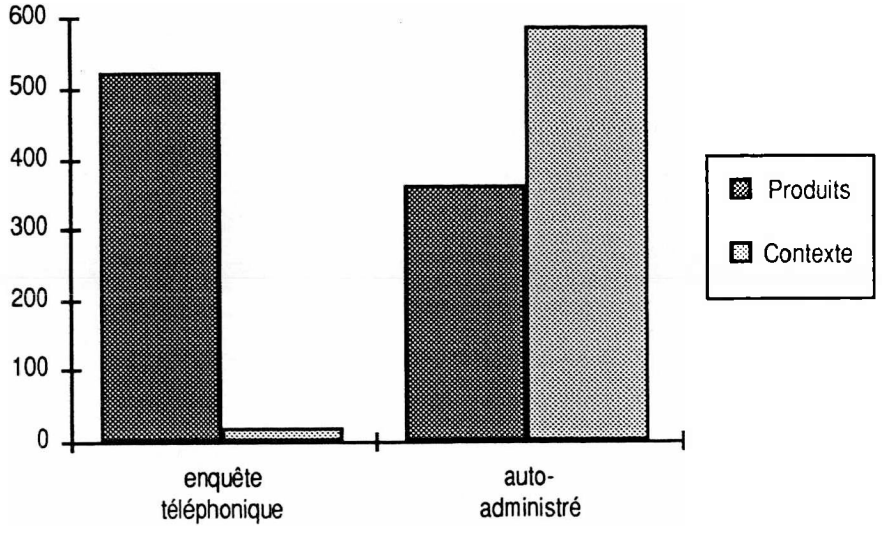
Reprenons pour simplifier non pas l'examen de toutes les classes, mais celui des deux classes qui se distinguent au début de l'analyse : celle des produits et celle du contexte.

Tableau 4 : Répartition des réponses suivant le sexe et le mode d'enquête

	Enquête téléphonique			Enquête auto-administrée	Ensemble
	Hommes	Femmes	Total	Femmes	
Réponses produits	94 %	97 %	95 %	38 %	67 %
Réponses contexte	6 %	3 %	5 %	62 %	33 %
Total	100	100	100	100	100

Si l'on considère les réponses téléphoniques, on remarque que la part des hommes qui ont répondu en envisageant le contexte (6 %) est plus forte que celle des femmes (3 %). Si l'on compare le comportement des femmes dans les deux échantillons, la différence est encore plus flagrante.

Graphique 8 : Les réponses des femmes selon le mode de questionnement



La différence de profil ne peut donc s'expliquer par la différence de structure de l'échantillon (absence d'hommes dans le panel), mais bien par l'incidence du mode de questionnement.

En bref, il semble que l'enquête téléphonique favorise des réponses purement descriptives centrées sur l'objet tandis que le questionnaire auto-administré autorise un élargissement du champ des associations.

Conclusion

La situation dans laquelle est placée la personne interrogée par téléphone (urgence, présence d'un interlocuteur) ont donc les effets suivants sur les réponses :

- absence de structure syntaxique ;
- réponses plus courtes ;
- vocabulaire moins riche ;
- réponses descriptives en termes de noms de produits ;
- efficacité pragmatique.

Il semble bien que la durée soit une condition indispensable à l'élaboration d'une réponse moins descriptive et plus conceptuelle, et sans doute plus personnelle.
La rémunération et la fidélisation des panélistes accroît aussi leur motivation, leur envie de "bien faire".

Le mode de questionnement a donc une forte incidence sur le type de réponses obtenues. Le choix d'un mode de questionnement doit tenir compte du type de résultat que l'on cherche à obtenir. Si l'on veut accéder aux représentations immédiates et spontanées, l'enquête téléphonique paraîtra bien adaptée. Si en revanche on cherche à obtenir l'expression la plus variée et complexe d'une représentation mentale, l'auto-administré sera préférable.

2. RÉDUCTION DE LA TAILLE DU VOCABULAIRE : ANALYSE SYNTAXIQUE, RÉSEAU SÉMANTIQUE

Dans cette partie, on cherchera à mesurer l'incidence de différents types de réduction de la taille du vocabulaire sur les résultats de l'analyse.

Nous disposons actuellement au CREDOC de trois types de méthodes pour réduire la taille du vocabulaire :
- la lemmatisation[6] et la reconnaissance des mots-outils à la façon ALCESTE ;
- la lemmatisation linguistique fondée sur une analyse syntaxique. Pour cela on dispose d'un analyseur syntaxique installé sur nos spécifications par Langage Naturel S.A. ;
- la réduction sémantique en utilisant le réseau sémantique construit par Langage Naturel.

Lemmatiser consiste ici à regrouper des formes différentes d'un même lemme telles qu'elles apparaissent dans un texte sous une forme unique, de manière à définir des unités lexicales canoniques :
- ramener les formes verbales fléchies, les variantes orthographiques à une forme canonique (infinitif, masculin, singulier, ...) ;
- réduire à une forme unique et radicale les différents aspects (participe, adverbe, ...) d'une même racine.

Pour comparer les effets de chacun de ces modes de réduction, nous avons pris comme base de test l'ensemble des réponses (2200) à une question ouverte sur l'effet de serre [7] :

> "D'après vous, en quoi consiste l'effet de serre et quelles sont ses conséquences ?"

Ce corpus a été lemmatisé par les trois méthodes et on a ensuite effectué dans des conditions identiques une analyse lexicale : même seuil de fréquence, même plan d'analyse... (respect du fameux "toutes choses égales par ailleurs").

---

[6] En fait, une stématisation, car on réduit non seulement les flexions, mais aussi les suffixages.

[7] Ce corpus nous a été fourni par le département GRETS d'EDF pour faire une étude comparative des logiciels SPAD.T et Alceste. L'étude complète figure dans l'Annexe 1.

2.1. La lemmatisation

Avant de passer au détail de l'analyse, il semble bon de revenir au débat sur la lemmatisation qui a fortement marqué "l'histoire" de l'analyse statistique des données textuelles.

2.1.1. Les enjeux du débat sur la lemmatisation

Les partisans de la lemmatisation se sont fortement opposés aux non partisans dans les débuts de l'analyse statistique des données textuelles. Pour les premiers, il s'agissait de réduire la diversité du vocabulaire pour mieux mettre en évidence les proximités sémantiques : la différence entre l'emploi du nom ou de l'adjectif est pour eux, notamment dans les réponses aux questions ouvertes, négligeable au regard de la proximité des concepts sous-jacents.

Au contraire, les partisans de la non lemmatisation, comme Christian Baudelot dans son introduction au livre de Lebart et Salem[8], justifient le choix de la non lemmatisation dans la mesure où pour eux des formes grammaticales sont connotées socialement.

Avec plus de recul, il semble que le débat théorique ait été désamorcé par les nouveaux développements techniques qui permettent une bonne lemmatisation. Au début de l'analyse des données textuelles, on n'avait aucun moyen de pratiquer une lemmatisation sans erreurs. Elle reposait sur des règles (par exemple la suppression du "s" en fin de mot), sur des listes pour les verbes irréguliers et pour les mots-outils, et sur des algorithmes de reconnaissance de la racine comme celui que propose Max Reinert dans son logiciel ALCESTE.

Les exemples d'erreurs, nombreux, sont liés à la très fréquente homographie en français. Par exemple, dans le logiciel ALCESTE, *vue*, *revue* étaient rattachés aux verbes *voir*, *revoir* parce qu'ils étaient reconnus comme une forme conjuguée de verbes irréguliers alors que dans le contexte ils désignaient respectivement "ce qui est vu" et un magazine. Au fur et à mesure, Max Reinert a modifié dans ses programmes les erreurs de lemmatisation que nous lui avons signalées. Ainsi *revues* et *vues* n'étaient plus lemmatisés. Mais cette solution est insatisfaisante puisqu'il n'y a plus de distinction faite entre le nom et le participe passé. Seul le recours à des lexiques et à un analyseur syntaxique permet de résoudre les problèmes

[8] LEBART Ludovic, SALEM André, *Analyse statistique des données textuelles*, Dunod, Paris, 1988.

d'homographie et d'obtenir une lemmatisation linguistiquement correcte. Les entrées de dictionnaires représentent des exemples parfaits de lemmatisation. Seuls les verbes à l'infinitif, les noms au singulier et les adjectifs au masculin singulier constituent des entrées de dictionnaire.

Une bonne lemmatisation consiste à ramener à l'infinitif toutes les formes conjuguées, au singulier tous les noms et au masculin singulier les adjectifs. Il faut passer par l'analyse syntaxique de la phrase pour désambiguïser les homographes et pouvoir rattacher le terme à sa forme lemmatisée.
Dans l'exemple suivant :

```
Il trouve l'information dans les revues littéraires.
```

seule une analyse syntaxique permet de rattacher *revues* à *revue* et non pas à *revoir*. Cette étape est de plus indispensable si l'on veut par la suite se connecter à des réseaux sémantiques construits à partir de dictionnaires.
Présentée de cette manière, l'opération paraît simple, bien délimitée. Mais dès que l'on passe à la réalisation, apparaissent des cas complexes comme le traitement de la négation, du passif, des participes passés : les solutions ne se font qu'à coup de conventions et de compromis.

2.1.2. Lemmatisation et analyse syntaxique

Rappelons que dans un des premiers logiciels d'analyse des données textuelles, SPAD.T, il n'y avait pas de lemmatiseur, et que pour réduire la taille du vocabulaire on éliminait les mots apparaissant moins de n fois ou ayant moins de n lettres (moyen très rustique pour éliminer une partie des mots grammaticaux). La méthode de lemmatisation ALCESTE et l'analyseur syntaxique ont apporté de remarquables améliorations par rapport à ce système ignorant du fonctionnement linguistique.

2.1.2.1. La "lemmatisation" ALCESTE

On emploie le terme lemmatisation dans un sens plus large puisqu'ici il englobe aussi la reconnaissance et l'élimination des mots-outils.

Une première étape consiste à reconnaître les mots-outils grâce à différents dictionnaires inclus dans le logiciel ALCESTE. Il s'agit des articles, des prépositions, des conjonctions, des pronoms (cette liste contient 321 termes ou locutions) et des auxiliaires *être* et *avoir*. Ces mots peuvent servir de variables illustratives.

Dans un second temps, les formes graphiques simples autres que les mots-outils sont regroupées. Deux méthodes de regroupement sont utilisées :

- la première permet de ramener à l'infinitif toutes les formes conjuguées des verbes irréguliers grâce à un dictionnaire.
Par exemple, les formes simples suivantes "*connu*", "*connaissant*", "*connais*", "*connaîtrons*", "*connaître*", "*connaîtront*", sont ramenées à la forme réduite "*connaître*" grâce à ce premier regroupement.

- la seconde méthode consiste à regrouper les formes du corpus associables à une même racine. Pour être réduite à sa racine, la forme associée doit se composer de cette racine et d'une désinence reconnue. Cette racine servira de support à l'identificateur de sens.
Par exemple, les formes suivantes "*augmentant*", "*augmentation*", "*augmenté*", "*augmentations*", "*augmente*", "*augmentent*" ont été ramenées à la forme réduite "augment+".

Les différents types de réduction effectués par ALCESTE sont reconnaissables aux signes terminant les formes :
- les formes réduites par reconnaissance de la racine et de la désinence se terminent par le symbole " . " (principalement pour les verbes irréguliers).
- les formes réduites par simple reconnaissance de la racine se terminent par le symbole " < ".
- les formes réduites par reconnaissance de la désinence, dans le cas où plusieurs formes du corpus commencent par la même racine, se terminent par le symbole " + ".

Mais ce système de lemmatisation hétérogène (listes et règles) engendre des erreurs qu'il n'est pas toujours facile d'identifier à moins de reprendre un par un tous les termes. Par exemple, *fraise*, est ramené à *frais+* comme *frais*, mais *fraîche* n'est pas ramené à *frais+*.[9]

[9] Cette erreur a été corrigée dans les versions ultérieures du logiciel.

2.1.2.2. La lemmatisation syntaxique

Afin d'obtenir une lemmatisation fiable, on dispose en amont de l'analyse lexicale d'un analyseur syntaxique qui, en binôme avec un lexique électronique, lemmatise le texte d'origine. L'analyseur syntaxique a été mis au point par Patrick Constant de Langage Naturel S.A. et a été installé d'après nos spécifications.

Patrick Constant dans sa thèse (Constant, 1991) montre clairement que le traitement automatique de la syntaxe - la mise au point d'une syntaxe analytique - entraîne des difficultés d'une autre nature que celles issues de l'analyse syntaxique traditionnelle ou analyse descriptive.

En grammaire traditionnelle, la description de la structure syntaxique d'une phrase est postérieure à sa compréhension, alors que dans une analyse automatique de la syntaxe, la compréhension n'intervient pas forcément. Le plus souvent, l'analyse sémantique est ultime. Il semble en effet exister un cheminement linéaire dans la démarche adoptée par la linguistique informatique qui va du lexique, à la syntaxe pour arriver à l'analyse sémantique puis pragmatique, sans doute parce que ces dernières sont les plus difficiles à formaliser.

Encore une fois, l'ordonnancement hiérarchique des différentes composantes correspond davantage à des impératifs méthodologiques qu'à une "vérité ontologique".
En effet, on ne sait que peu de choses sur le fonctionnement du processus de compréhension des énoncés. Si en ce qui concerne la pragmatique, il paraît évident que la phase de ré-interprétation perlocutoire est postérieure à celle de l'interprétation, en revanche on ne sait si la compréhension s'appuie d'abord sur le lexique ou sur la syntaxe, ou si le sens est compris à partir d'une prise en compte simultanée des diverses composantes. Mais que signifie cette simultanéité ?

Les problèmes majeurs que pose la syntaxe analytique sont ceux de l'attachement et de l'ambiguïté. L'attachement est la compréhension des liens entre les divers éléments. L'ambiguïté est le propre de la syntaxe analytique puisqu'en syntaxe descriptive elle est résolue par la compréhension du contexte.
En toute rigueur, la réduction est impossible, car le sens d'un mot est donné par le contexte et varie, dans le détail, infiniment.
Prenons un exemple :

> Ma soeur est venue à la maison avec un ami que je ne connaissais pas. J'en ai profité pour lui rendre sa vieille collection de timbres.

Lui renvoie évidemment à *ma soeur*. Mais cette anaphore ne peut être résolue que par une connaissance du monde externe à la phrase, et même au texte. Une bonne réduction devrait pouvoir remplacer *lui* par *soeur*.

Enfin, comme nous l'avons vu, tout type de lemmatisation implique certains choix, plus ou moins arbitraires. Voici les nôtres :

- les participes présents et passés ne sont pas ramenés à l'infinitif : respect des fonctions adjectives de ces formes conjuguées ;

- les verbes à la forme passive sont ramenés à l'infinitif et précédés d'un p_ :

Le soleil frappe directement sur la matière sans être filtré - les rayons ultraviolets détruisent tout.	Le soleil frapper directement sur le matière sans p_filtrer - le rayon ultraviolet détruire tout.

- les verbes sous la forme négative sont ramenés à l'infinitif et affixés par un n_ :

Le gaz carbonique et la pollution font écran entre le soleil et la terre où nous vivons : les transformations chimiques naturelles ne peuvent donc plus se faire.	Le gaz carbonique et le pollution faire écran entre le soleil et le terre où nous vivre : le transformation chimique naturel n_pouvoir donc plus se faire.

Remarquons que le traitement de la négation en analyse automatique du langage pose de nombreuses difficultés :

- selon que l'on valorise le choix d'un item lexical ou le fait qu'il soit nié. C'est cette deuxième option que l'on a choisie. Dans le cas inverse, on aurait :

le n_transformation chimique naturel pouvoir plus se faire.

- selon l'intensité du forclusif (*guère*, *pas*, *jamais*), le sens varie ;

- les locutions reconnues à l'aide d'un dictionnaire de 20 000 locutions sont marquées par un "blanc souligné" ("underscore") :

J'ignore ce que signifie l'effet de serre	Je ignorer ce que signifier le effet_de_serre

• on utilise pour l'analyse lexicale un fichier dans lequel sont éliminés tous les mots grammaticaux. Ce choix est légitime dans la mesure où l'analyse statistique est indifférente à la syntaxe.

La pollution reste en circuit fermé plus échauffement de l'atmosphère, de l'eau, tout cela entraîne la disparition et la dégradation de la vie animale et végétale.	Pollution rester circuit fermé réchauffement atmosphère eau entraîner disparition dégradation vie animal végétal

Une difficulté doit être signalée : la catégorie grammaticale des locutions n'est pas prise en compte : toutes les locutions sont conservées même des locutions grammaticales comme *autour de*, *même si*... Cela est assez incohérent avec le choix d'éliminer les mots grammaticaux et sera sans doute amélioré dans d'ultérieures versions. Pour le moment, on peut modifier manuellement le statut de ces locutions.

2.1.3. Incidence des différentes formes de lemmatisation sur l'analyse lexicale

Comme nous l'annoncions en introduction, nous allons voir l'effet de ces deux types de lemmatisation sur l'analyse du corpus "effet de serre".

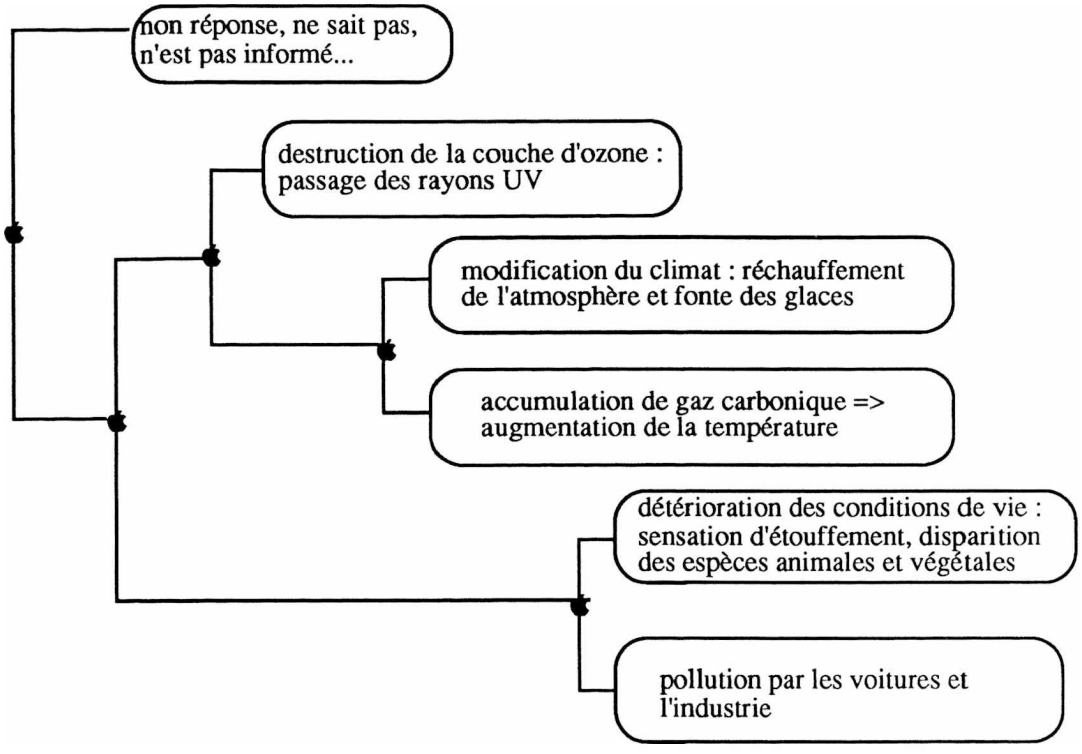
- première analyse : lemmatisation ALCESTE + mots-outils traités en variables illustratives.

- seconde analyse : lemmatisation linguistique avec analyseur syntaxique et lexique + élimination des mots grammaticaux.

Ensuite, dans les deux cas on éliminera les mots de fréquence inférieure à quatre et on effectuera une classification descendante hiérarchique dans les conditions habituelles, en retenant comme unité d'analyse la réponse.

On retrouve la même structure de l'arbre de classification avec une lemmatisation de type ALCESTE comme avec une lemmatisation "linguistique".

Graphique 9 : Typologie des réponses à la question sur l'effet de serre



Les différences les plus notables apparaissent lors de découpages plus fins mais c'est surtout la taille de chacune des classes qui varie.

Tableau 5 : Tableau récapitulatif des classifications obtenues selon le mode de lemmatisation

Nom des classes	Lemmatisation ALCESTE	Lemmatisation linguistique analyse syntaxique et lexique
Destruction de la couche d'ozone	15 %	14 %
Accumulation de gaz carbonique	22 %	22 %
Modification du climat : réchauffement de l'atmosphère et fonte des glaces	26 %	27 %
Pollution (voiture et industrie)	12 %	10 %
Détérioration des conditions de vie (étouffement, disparition des espèces)	16 %	18 %
Ne sait pas, pas assez informé, aucune idée	4 %	1 %
U C E non classées	5 %	8 %

Compte tenu de la fiabilité des classes, une différence de quelques pour cent est parfaitement négligeable.

Le matériau textuel constitué par les réponses à une question ouverte est tellement redondant que des erreurs de lemmatisation n'ont finalement qu'une incidence minime. Le fait que la lemmatisation soit correcte accroît en tout cas la lisibilité des résultats.

Notons enfin que l'analyseur syntaxique spécialement conçu pour l'analyse lexicale commence à être utilisé dans d'autres domaines. Dans la base de données de l'Observatoire des Consommations Alimentaires, il devrait faciliter les recherches sur les libellés en langage naturel et le codage des produits alimentaires. Une maquette de traitement est en cours d'élaboration et devrait être implémentée dans les mois à venir[10].

2.2. La reconnaissance des locutions

2.2.1. Rôle des locutions en analyse lexicale

Dans le débat sur la lemmatisation, le problème de la reconnaissance des locutions a été assez peu souvent abordé. Certains linguistes se sont penchés sur la question mais l'interconnexion avec la statistique lexicale a été rarement faite. On présentera rapidement les trois grands types de démarches qui permettent de repérer automatiquement les noms composés dans les corpus, ensuite, on cherchera à montrer les moyens utilisés en analyse lexicale au CREDOC pour repérer les noms composés.

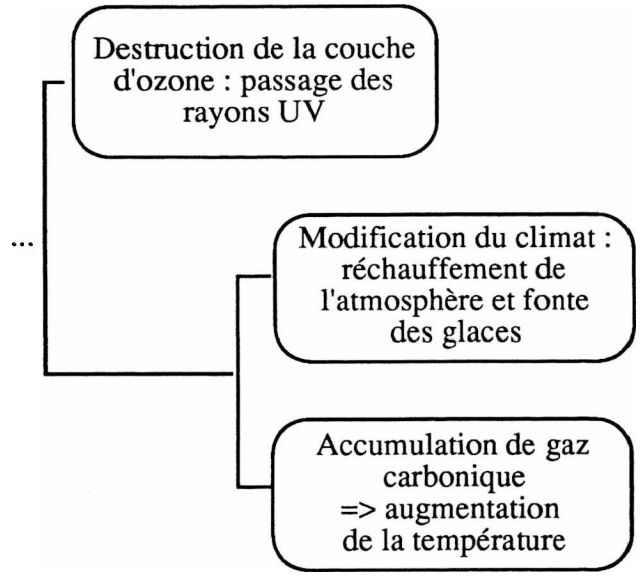
La reconnaissance des noms composés est une étape cruciale en analyse lexicale pour deux raisons :

- c'est un moyen pour réduire la taille du vocabulaire ;
- c'est un moyen d'éviter des cooccurrences systématiques qui ont tendance à atténuer la diversité du reste du corpus et surtout à modifier le poids des classes. En effet, quand une locution n'est pas reconnue, il y a une cooccurrence systématique de deux ou plusieurs termes qui a tendance à créer un pôle qui "attire" des réponses qui pourraient être rattachées à d'autres classes.

[10] Rapport CREDOC "Etat d'avancement de la base de données de l'Observatoire au 30 mai 1993" (confidentiel).

Prenons comme exemple, les réponses sur l'effet de serre pour examiner l'incidence sur l'analyse de la reconnaissance de la locution *couche d'ozone*. Nous ne reprenons que la partie de l'arbre de classification où sont intervenues des modifications.

Graphique 10 : Incidence de la reconnaissance des locutions sur l'analyse lexicale



Locution "couche d'ozone"	
Pas de reconnaissance	**Reconnaissance**
29 %	22 %
37 %	43 %
34 %	35 %

On voit clairement sur le graphique ci-dessus que la taille des classes varie considérablement : la classe *couche d'ozone* perd sept points au bénéfice de la classe *modification du climat*. Cette analyse comparée est à la fois inquiétante et rassurante : inquiétante parce que la taille de chacune des classes varie de manière non négligeable, rassurante parce que la structure globale (l'arbre de classification) est identique dans les deux cas. On peut faire confiance à la structure d'ensemble de la représentation qui est assez peu sensible aux éléments micro-structuraux. Mais, il est nécessaire d'être toujours prudent dans l'analyse des pourcentages, car une simple locution non reconnue peut faire varier le classement des réponses. D'où l'importance d'une bonne reconnaissance des locutions.

2.2.2. Reconnaissance automatique des locutions

La reconnaissance des locutions ou noms composés est difficile de manière automatique puisque la plupart du temps il n'y a pas de signe typographique (tiret ou autre) entre les différents termes d'un mot composé.

Une journée sur les noms composés a été organisée par Pierre Cadiot (Université de Paris VIII), Benoît Habert (ENS de Fontenay) et Christian Jacquemin (Université de Nantes) au mois de juin 1992 (Cadiot, Habert, Jacquemin, 1992). Dans le compte-rendu de cette journée, sont présentés trois grands types de traitement pour les noms composés :

- l'approche statistique : la mise en évidence de couples ou segments répétés (séquence de mots qui reviennent un certain nombre de fois) permet de reconnaître les locutions dans un corpus donné. C'est un bon moyen pour identifier de nouveaux mots composés qui entrent dans l'usage.

- les approches structurelles : elles considèrent que l'étude sur les noms composés se situent au croisement des divers domaines linguistiques : lexique, syntaxe, sémantique. La solution la plus généralement adoptée est la constitution de listes électroniques les plus exhaustives possible qui prennent en compte toutes les segmentations.

- les approches précédentes sont insatisfaisantes dans la mesure où elles ne prennent pas en compte toutes les variations possibles sur les noms composés. Christian Jacquemin, au début à l'aide d'un système de règles grammaticales, aujourd'hui à l'aide des réseaux neuronaux, propose un système de détection robuste des noms composés à travers leurs variations (Jacquemin, 1992 a et b).

Dans le cadre de ce débat, il faut d'abord noter que la définition de l'objet "nom composé" est particulièrement floue et varie selon les types d'approches. Ne serait-il pas temps d'essayer de délimiter le champ d'étude de manière plus rigoureuse, c'est-à-dire de proposer une définition qui englobe la notion de variation, soit une définition analytique, soit une définition descriptive qui recense toutes les catégories, soit une définition pragmatique liée à la manière dont ils sont repérés ? Cette dernière approche est la moins satisfaisante à un niveau théorique mais c'est pour le moment la plus efficace au niveau pratique.

2.2.3. La méthode adoptée au CREDOC

Pour l'analyse des données textuelles, nous adoptons une démarche qui combine les deux premières méthodes de repérage automatique vues ci-dessus : listes électroniques et analyse statistique.

2.2.3.1. Les listes électroniques

Le logiciel ALCESTE dispose d'une liste de 109 locutions qui sont avant tout des locutions grammaticales du type "à condition", "bien qu+", "ce qu+", "de manière", "de même", "en dépit", "en effet".
Dans le lexique fourni par Langage Naturel SA, qui comprend 60 000 entrées, on a une liste de 20 000 locutions. Ainsi, la première étape avant l'analyse consiste à repérer dans le texte les locutions répertoriées dans les listes, et à relier les différents termes constituant le nom composé par un signe typographique particulier : le tiret ou le blanc souligné pour qu'il soit par la suite analysé comme un seul mot plein.
L'utilisation d'une liste n'est pas suffisante pour repérer tous les noms composés, parce qu'ils ne sont pas tous répertoriés. Ceci est dû à la définition assez floue des noms composés et au fait que c'est un moyen très fréquemment utilisé pour créer de nouveaux termes (surtout dans les domaines techniques). La mise à jour de ces listes devrait être faite très régulièrement.

Ainsi dans notre corpus sur l'effet de serre, les noms composés *gaz carbonique* et *couche d'ozone* n'apparaissaient pas dans les listes contrairement à : *effet de serre*, *rayons ultraviolets* ...

2.2.3.2. Les couples et segments répétés

Il faut donc disposer d'un autre moyen pour les repérer. C'est grâce à l'analyse des couples et segments répétés que l'on sera sûr d'avoir bien identifié tous les noms composés apparaissant dans le corpus.

L'analyse des couples et segments répétés consiste à repérer les séquences de deux à cinq mots qui apparaissent un certain nombre de fois dans les textes. Ensuite les séquences ainsi identifiées seront ajoutées au dictionnaire des locutions. On a ainsi un repérage des nouvelles locutions pertinentes pour le texte analysé. C'est en ayant recours à cette méthode, c'est-à-dire en constituant le dictionnaire des couples et segments répétés et en observant ceux qui ont la fréquence la plus élevée, qu'on a repéré les locutions *gaz carbonique* et *couche d'ozone*.

L'utilisation simultanée de listes électroniques et de critères statistiques est une solution pragmatique tout à fait satisfaisante pour l'analyse lexicale.

2.3. La réduction sémantique

Les points que nous venons d'examiner (lemmatisation, reconnaissance des locutions) agissent sur la structure morpho-syntaxique mais ne concernent pas la sémantique. Or est apparue au CREDOC (Yvon, 1990 ; Lahlou, 1992) la nécessité en analyse lexicale d'avoir recours à des outils de sémantique (repérage des synonymes, des mots appartenant à la même famille...) pour faire du classement sur des signifiés (sens) et non sur des signifiants (formes graphiques). Comment reconnaître automatiquement que deux mots distincts sont synonymes, que deux formes graphiques identiques ne renvoient pas au même sens (homographie, polysémie) ? C'est pour ces raisons que nous nous sommes intéressés aux réseaux et que nous en avons fait installer un dans le département Prospective de la Consommation par Langage Naturel S.A. .

Nous présentons d'abord les dernières méthodes utilisées pour construire des réseaux sémantiques (à partir de dictionnaires ou de corpus de textes) avant de décrire les fonctionnalités du réseau Langage Naturel et du système de réduction SIT adapté à l'analyse lexicale. Nous chercherons enfin à mesurer l'impact de ce type de réduction sur les résultats de l'analyse lexicale.

2.3.1. Lexiques informatisés : vers des réseaux sémantiques

Il existe deux types d'approches des lexiques informatiques :

- l'une en terme de listes ou de thesaurus qui consiste à énumérer tous les termes constitutifs du lexique assortis d'informations : description morpho-syntaxique, étymologie, définition, synonymes, exemples, ... C'est le principe même des dictionnaires électroniques. Cette approche est constructive dans la mesure où elle cherche à expliciter au mieux le concept.

- l'autre en terme de réseau qui cherche à structurer le lexique en mettant en évidence des relations entre les termes. Les relations peuvent être de nombreux types, relation de synonymie, d'hyponymie/hyperonymie, de méréologie (relation du tout à la partie), relation morpho-syntaxique (appartenance à une même famille lexicale). Cette approche est

différentielle en ce sens que le concept n'est pas défini en lui-même mais par les relations sémantiques qu'il entretient avec les autres termes.

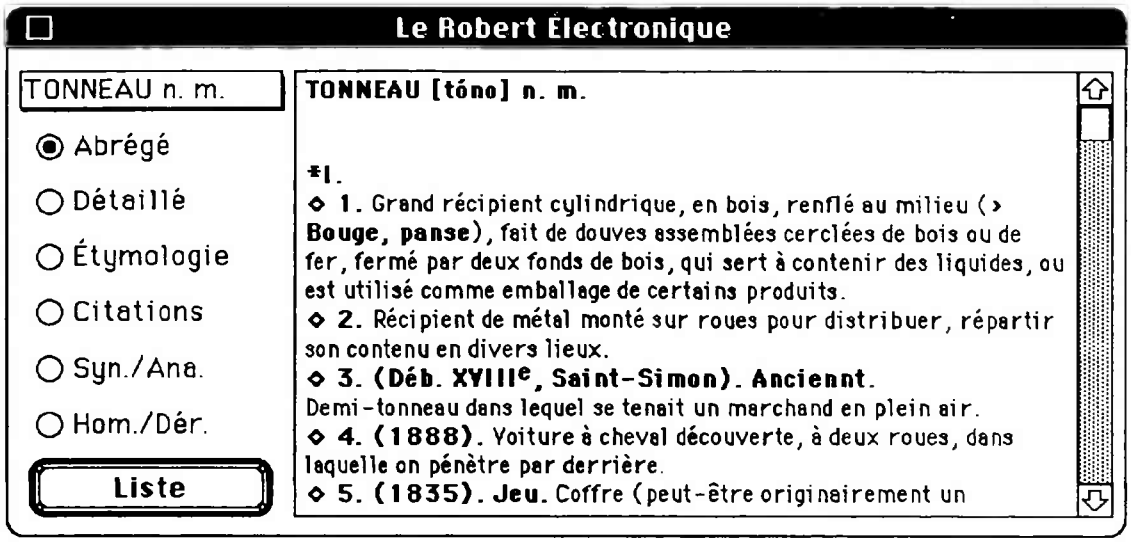
La seconde approche, en terme de réseau, semble être actuellement la plus productive pour le traitement automatique du sens. Les linguistes s'appuient sur deux types de matériaux pour construire des réseaux sémantiques : les dictionnaires et les corpus de textes.

2.3.1.1. Dictionnaires et réseaux sémantiques

Nous nous appuyons sur les travaux de Miller et Fellbaum (1991) qui semblent représenter une voie d'avenir pour l'implémentation de "bons" réseaux sémantiques.
Comme point de départ à la sémantique lexicale, Miller et Fellbaum (1991) traitent le mot comme une association arbitraire entre un signifiant (word form) et un signifié (word meaning) se rattachant ainsi à la tradition saussurienne. Par la dissociation entre forme et signifié, les problèmes de la synonymie et de la polysémie peuvent être en principe résolus.
Comment représenter, en pratique, le signifiant et le signifié d'un terme ? La forme graphique est un bon représentant du signifiant. Pour le moment, les définitions de dictionnaires semblent être les meilleurs outils pour représenter le signifié.

> "Definitions can play the same role in simulation that meanings play in the mind of a language user"
> (Miller, Fellbaum, 1991).

Les dictionnaires ont été les premiers outils utilisés pour la constitution manuelle ou automatique des réseaux sémantiques. Dans la structure de la définition lexicographique, on trouve toutes les informations qui permettent de constituer des réseaux de termes. La réalisation de dictionnaires électroniques a contribué à structurer la définition de manière plus précise. Prenons à titre d'exemple le Robert électronique : chaque définition est organisée selon divers champs qui permettent une consultation focalisée.



La rubrique "synonymes / analogues" permet de créer immédiatement des liens de synonymie entre le terme défini et les synonymes.

TONNEAU
définition
◊ 1. Grand récipient cylindrique, en bois, renflé au milieu (> Bouge, panse)
synonymes et analogues
> Barrique, botte, boucaut, feuillette, fût, futaille, muid, pièce, poinçon.
> Foudre, tonne.
> Baril, quartaut, tonnelet.
> Bonde, cannelle, cerceau, cercle, chantepleure, douelle, douve, enfonçure, fonçailles, fond, jable, robinet, sommier, traversin.
> Feuillard, merrain.
> Tonnelier.
> Poinçon, rouanne.
> Entonner; entonnage.
> Haquet.
> Chape.
> Bondon, broche, douzil, fausset.
> Tire-bonde.
> Baquetures, coulage.
> Ouiller, rembouger.
> Méchage, mécher.
> Baissière.
> Ivrogne.

Malheureusement dans le Robert électronique cette rubrique "synonymes / analogues" regroupe des termes qui entretiennent avec le terme défini différents types de relation : synonymie pour [tonneau, barrique], méréonymie (relation de la partie au tout) pour [cerceau, tonneau], ...

Un réseau sémantique construit automatiquement à partir du dictionnaire Le Robert hériterait de la précision insuffisante de la rubrique "synonymes / analogues".

Enfin, le contenu même de la définition permet de repérer des relations d'hyponymie / hyperonymie. La structure habituelle de la définition d'un terme est :

```
mot défini :    genre prochain + traits spécifiques
                archisémème + sèmes
```

La relation entre le mot défini et le genre prochain définit souvent une relation hyponymie / hyperonymie.

1. **FRAISE** [fRèz] n. f.
◊ 1. **Fruit** du fraisier; masse charnue comestible, rouge, qui en fait partie.

Fruit est un hyperonyme de **fraise.** C'est en cela que l'on peut dire que le contenu des définitions lexicographiques permet de construire une organisation hiérarchisée du lexique.

Toutes ces informations permettent de reconstituer des liens (graphes en étoile) autour du terme défini. L'unification de tous ces graphes permet de reconstituer un réseau sémantique global.

La qualité d'un réseau construit de manière automatique à partir d'un dictionnaire dépend de la cohérence des définitions. Il est préférable que :
- tous les termes appartenant à une même famille soient raccordés à un même hyperonyme ;
- la relation de synonymie soit symétrique et transitive ;
- les types de relations soient clairement identifiés et distingués.

Il existe déjà sur le marché des réseaux sémantiques construits automatiquement à partir de dictionnaires. Par exemple, Langage Naturel S.A. a construit à partir du Larousse électronique son réseau.

Miller et son équipe ont construit manuellement un thesaurus sous forme de réseau (Wordnet) dont la structure est assez novatrice. Il constitue des ensembles de synonymes (synsets) qui regroupent toutes les occurrences d'un concept. Dans Wordnet, les 34 000 mots sont regroupés dans 26 000 ensembles de synonymes, qui représentent les signifiés. Une fois établie la représentation des signifiés dans le thesaurus, le problème central est l'identification des relations sémantiques entre les différents signifiés ou ensembles de synonymes.
L'intérêt de la démarche de Miller et Fellbaum (1991) est d'avoir défini pour chaque grande catégorie syntaxique (noms, adjectif, verbe) des relations sémantiques pertinentes. En effet, s'appuyant sur des expériences psychologiques, les auteurs montrent que selon les catégories morpho-lexicales, ce ne sont pas les mêmes relations qui structurent le lexique. En résumé, pour le nom, les relations pertinentes sont la synonymie, l'hyponymie, l'antonymie, la

méréonymie ; pour l'adjectif, la synonymie, l'antonymie et la gradation ; pour le verbe, la troponymie (V1 c'est V2 "d'une certaine manière"), l'opposition, la causalité et l'implication. Le danger d'une telle approche est de ne pas établir de liens sémantiques entre les grandes catégories. Les auteurs évitent cet écueil en montrant comment on circule de la catégorie du nom à celle du verbe et de l'adjectif.

2.3.1.2. Corpus de textes et relations sémantiques

Nous venons de voir que les définitions de dictionnaire peuvent être et sont déjà largement utilisées pour représenter le sens des termes et, plus particulièrement dans une approche différentielle, pour mettre en évidence des relations entre termes. Aujourd'hui, les travaux en sémantique lexicale commencent à s'appuyer sur des corpus de textes et non plus seulement sur les définitions de dictionnaires. La multiplication des corpus sur support informatique a sans doute joué un rôle important dans cette évolution. Les corpus de textes deviennent des sources d'information sur le langage, et plus précisément sur l'organisation du lexique.

Hearst Marti A. (1992) propose une méthode pour repérer de manière automatique des relations d'hyponymie/hyperonymie sur de larges corpus de textes. L'objectif est de rechercher toutes les structures lexico-syntaxiques qui peuvent signaler ce type de relations. En anglais, il montre que six structures peuvent introduire des relations d'hyponymie/hyperonymie. La première est :

> *NP such as {NP', NP", ...(and/or)} NPn*
> ex : The bow lute, such as the Bambara ndang, is plucked...
> Bow lute est l'hyperonyme de Barbara ndang

Pour identifier de nouvelles structures grammaticales qui permettent de repérer des couples de termes en liaison hyponymique, Hearst (1992) propose la démarche suivante :

> 1. choisir un type de relation sémantique : synonymie, hyponymie...
> 2. choisir une liste de couples de termes qui vérifient cette relation. Ceux-ci peuvent être repérés automatiquement à l'aide de structures identifiées au préalable.
> 3. rechercher tous les contextes dans lesquels ces couples de termes cooccurrent ;
> 4. identifier la structure lexico-syntaxique qui introduit la relation sémantique ;
> 5. repérer tous les autres couples ou ensembles de termes mis en relation par cette structure.
> L'objectif est bien de repérer des couples de termes qui sont reliés par une relation d'hyponymie.

L'identification des modèles lexico-syntaxiques attestés qui indiquent certains types de relations entre termes permet sans doute la constitution de réseaux sémantiques plus

complexes, qui dépassent la science du et s'approchent davantage de l'usage de la langue et surtout de sa capacité à innover.

Concrètement l'identification des relations d'hyponymie permet de vérifier et surtout d'augmenter les lexiques informatisés construits à partir de dictionnaires. C'est l'objet de l'article de Hearst (1992) de montrer comment l'identification de relations d'hyponymie sur des corpus de textes permet d'enrichir les lexiques existants et en particulier le lexique Wordnet dans lequel, comme nous l'avons vu, les termes sont organisés en ensembles de synonymes (synsets) et reliés entre autres par des relations d'hyponymie / hyperonymie.

Cette méthode permet aussi d'appréhender le sens d'un mot inconnu par son hyperonyme ou synonyme et donc d'enrichir les dictionnaires. Enfin, l'identification de formes lexico-syntaxiques désignant des relations sémantiques permet de relier non seulement des entités lexicales mais aussi des syntagmes nominaux. Cette méthode peut donc servir à enrichir les réseaux sémantiques mais également à constituer des réseaux de syntagmes nominaux particulièrement profitables dans des domaines spécialisés.

Pour le moment, Hearst (1992) ne propose que la détection de couples d'hyponymes / hyperonymes. Sa méthode semble avoir échoué pour le repérage de liens sémantiques comme la méréonymie. De plus, sa démarche n'est valable que pour les noms et les syntagmes nominaux. Les réflexions de Miller et Fellbaum (1991) nous conduisent à penser que pour les noms et adjectifs, les structures introduisant des relations sémantiques seront sensiblement différentes.

Cette approche et la procédure proposée pour identifier de nouvelles structures de relations sémantiques ont un intérêt évident, mais elles paraissent très coûteuses en temps et difficilement implémentables (l'algorithme proposé n'a pas encore été programmé). La formalisation automatique de la structure (étape 4) en analysant les contextes de cooccurrences est loin d'être évidente. En revanche, le repérage des structures syntaxico-lexicales, une fois qu'elles sont formalisées, ne pose pas de grandes difficultés puisque seules des règles de grammaire locales sont nécessaires pour identifier les structures lexico-syntaxiques. Une difficulté supplémentaire vient du fait que la reconnaissance de ces structures sur de vastes corpus de textes introduit beaucoup de bruit. Sur un corpus encyclopédique (*Grolier's American Academic Encyclopedia*), Hearst (1992) a trouvé 7067 phrases qui contenaient l'expression *such as*. Seules 152 relations d'hyperonymie ont été identifiées dans tous ces contextes. En clair, la structure introduit quelques relations d'hyperonymie qui peuvent être récupérées dans un réseau sémantique mais, à 98 %, elle ne signale pas des relations pertinentes du point de vue du réseau, le plus souvent parce qu'il s'agit de groupes nominaux ou de termes sous des formes non canoniques qui n'apparaissent pas dans les dictionnaires.

Enfin, si l'on s'attache au détail des relations trouvées, on peut noter quelques phénomènes assez peu satisfaisants :
- les relations mises en évidences sont encore plus souvent que dans les dictionnaires dépendantes du contexte ;
- l'écart sémantique entre l'hyponyme et l'hyperonyme est variable. Trouver une relation entre *caniche* et *être vivant* n'a sans doute pas un grand intérêt.

S'il paraît intéressant en théorie d'utiliser les corpus de textes pour enrichir les réseaux sémantiques, dans la pratique le travail exigé est encore quelque peu démesuré. Les outils de lemmatisation (lexiques et analyseurs syntaxiques) devraient rendre les recherches plus productives. Une formulation plus souple des règles de grammaire devrait aussi rendre les recherches plus fructueuses.

Pour le moment au CREDOC nous disposons d'un arbre sémantique construit à partir d'un dictionnaire électronique par Langage Naturel S. A. Dans les années à venir, nous serons sans doute amenés à utiliser des réseaux plus complexes. Par exemple, pour l'analyse automatique du sens des réponses à des questions ouvertes, il serait sans doute judicieux d'avoir recours à des réseaux construits à partir du discours oral.
Aujourd'hui, quel est pour nous l'intérêt de disposer d'un arbre sémantique ?

2.3.2. Objectifs et méthodes

L'objectif était de réduire la taille du vocabulaire en ayant recours à la sémantique et non plus simplement à la reconnaissance des formes grammaticales. Nous cherchons donc à parvenir à identifier des classes de synonymes ou de mots appartenant à la même famille morpho-syntaxique (du type *spacieux, espace...*) à l'intérieur du corpus et, dans ce cas seulement, remplacer tous les termes d'une classe soit par leur hyperonyme quand il existe (par exemple *gâteau, tarte* ... seront remplacés par *dessert*), soit par un des termes de la classe (celui qui aurait la fréquence d'apparition la plus élevée par exemple). La transformation du lexique n'a lieu que dans la mesure où elle permet de résumer l'information.
Si les termes *équilibré, mesuré, raisonnable, pondéré, sage, sain* apparaissent tous au moins une fois dans le corpus, chacune de leurs occurrences pourrait être ramenée à la forme *équilibré*.
Ainsi sans qu'il y ait une altération sensible du sens, les analyses lexicales devraient beaucoup gagner en stabilité.

Cette étape est postérieure à la lemmatisation syntaxique.

Pour qu'il puisse y avoir réduction sémantique, il faut :

- disposer d'un réseau sémantique fabriqué à partir d'un dictionnaire électronique ;
- analyser le contexte. En effet s'il ne s'agissait que de remplacer chaque terme par un de ses synonymes ou par son hyperonyme, on aurait simplement une translation lexicale sans aucune ou pratiquement aucune réduction de la taille du vocabulaire. Avant de réduire, l'analyse du contexte est indispensable. Ainsi, le système de réduction mis en place par Langage Naturel S.A. sur nos spécifications privilégie les réductions "utiles" : un terme qui dispose d'un synonyme ou d'un hyperonyme, n'est réduit que si au moins un autre terme du corpus admet ce même terme réducteur.

Dans le réseau sémantique, les mots sont reliés entre eux par quatre types de relation :

- relation d'hyperonymie "forte"

L'hyperonyme est "un mot (et notamment, nom) dont le sens (compréhension logique) inclut celui d'autres mots (noms)" (Le Grand Robert électronique).

soleil	hyp	astre
légume	hyp	végétal

- relation d'hyperonimie "souple"

bouleversement	ext	changement
élévation	ext	augmentation

- relation de synonymie

désastre	syn	malheur
débordement	syn	inondation
déluge	syn	inondation

- appartenance à la même famille morpho-syntaxique

atmosphérique	=>	atmosphère
accroissement	=>	croître

Ces types de liens fabriqués à partir des définitions lexicographiques ne sont pas toujours optimaux dans la mesure où ils sont le plus souvent valables dans un type de contexte et non dans tous. Exemple : “fraise” peut être un fruit, un vêtement, un instrument.

Dans ce réseau sémantique, il existe donc des relations qui peuvent surprendre et paraître arbitraires ; elles ne le sont pas, mais sont liées à des problèmes de polysémie et de polyusage.

Une autre des difficultés que pose la réduction sémantique est celle de la convergence/ divergence. Bien entendu comme nous venons de le voir, on ne transforme un terme en un autre que si un autre terme converge vers le même terme. Par exemple, *catastrophe* est transformé en *malheur* parce que *désastre* qui admet aussi *malheur* comme synonyme apparaît aussi dans le corpus.

catastrophe	syn	malheur
désastre	syn	malheur

Dans ce cas précis, la convergence ne pose aucune difficulté. En revanche, dans les exemples suivants, les difficultés sont plus redoutables :

élévation	ext	augmentation
surcroît	syn	augmentation
accroître	ext	augmenter
amplifier	ext	augmenter
croître	ext	augmenter
multiplier	syn	augmenter
accroissement	fam	croître
grandir	syn	croître

On a ici un cas qui montre clairement les difficultés posées par la réduction sémantique. Alors que *accroître* et *accroissement* auraient tous deux dû converger vers le même terme, ils divergent. Le problème vient de la construction de l'arbre sémantique lui-même.

Nous donnons ci-après des extraits du réseau sémantique. Le terme central est encadré. Autour apparaissent les mots qui sont liés au terme, ainsi que le type de relation qui assure le lien.

```
allongement      -(syn)->
agrandissement   -(syn)->
                          accroissement
                          -(syn)->       augmentation
                          -(fam)->       croître
```

Ce n'est pas dans le type de liens qui structurent le réseau qu'apparaissent les plus grosses difficultés, mais dans la manière d'analyser le corpus pour prendre des décisions de réduction. Il se pose un problème de priorités entre les différents types de liens qui ne peut être résolu que par une certaine anticipation des résultats, qu'un lecteur humain fait naturellement, mais qui est difficile à programmer en raison de sa dépendance au contexte.

Enfin, certaines transformations révèlent un non respect de la simple règle de convergence. On s'étonnera de voir que :

automobile	syn	voiture
voiture	syn	véhicule
automobile	ext	véhicule

Le mot *véhicule* n'est pas considéré comme un synonyme d'*automobile* mais comme un hyperonyme "souple".

Telle qu'elle est traitée dans le réseau, la synonymie n'est pas toujours transitive. Il se pourrait bien que, en dehors du contexte des dictionnaires, cette relation soit effectivement approximative, non transitive et dépendante du contexte.

Dans la constitution de l'arbre sémantique telle qu'elle a été faite par Langage Naturel S.A., on peut relever quelques insuffisances :

- la première est qu'elle est exclusivement fondée sur les dictionnaires alors que l'information contenue dans des ensembles de textes aurait pu l'enrichir considérablement ;
- la seconde est que les relations sémantiques retenues manquent parfois de précision. La relation de synonymie n'est pas transitive et la relation *extension* définie comme une forme d'hyponymie faible n'est pas suffisamment précise. Ensuite, on peut regretter que les relations n'aient pas été définies en fonction des catégories morpho-syntaxiques. L'approche de Miller est à ce propos sans doute plus pertinente ;
- enfin, construit de manière automatique à partir d'un dictionnaire, le réseau sémantique hérite des incohérences dans la constitution du dictionnaire.

La réduction sémantique opérée à l'aide de l'outil de Langage Naturel S. A. est actuellement insatisfaisante car la prise en compte du contexte est trop superficielle. En particulier, seule une analyse fine du contexte aurait permis de résoudre les problèmes de polysémie. Pour le moment, quand un terme admet plusieurs sens, il n'y a tout simplement pas de réduction.
La médiocrité des résultats actuels n'invalide pas en tout cas le recours à des réseaux sémantiques construits à partir de dictionnaires pour l'analyse du sens et plus particulièrement pour la réduction sémantique, mais du travail reste à faire.

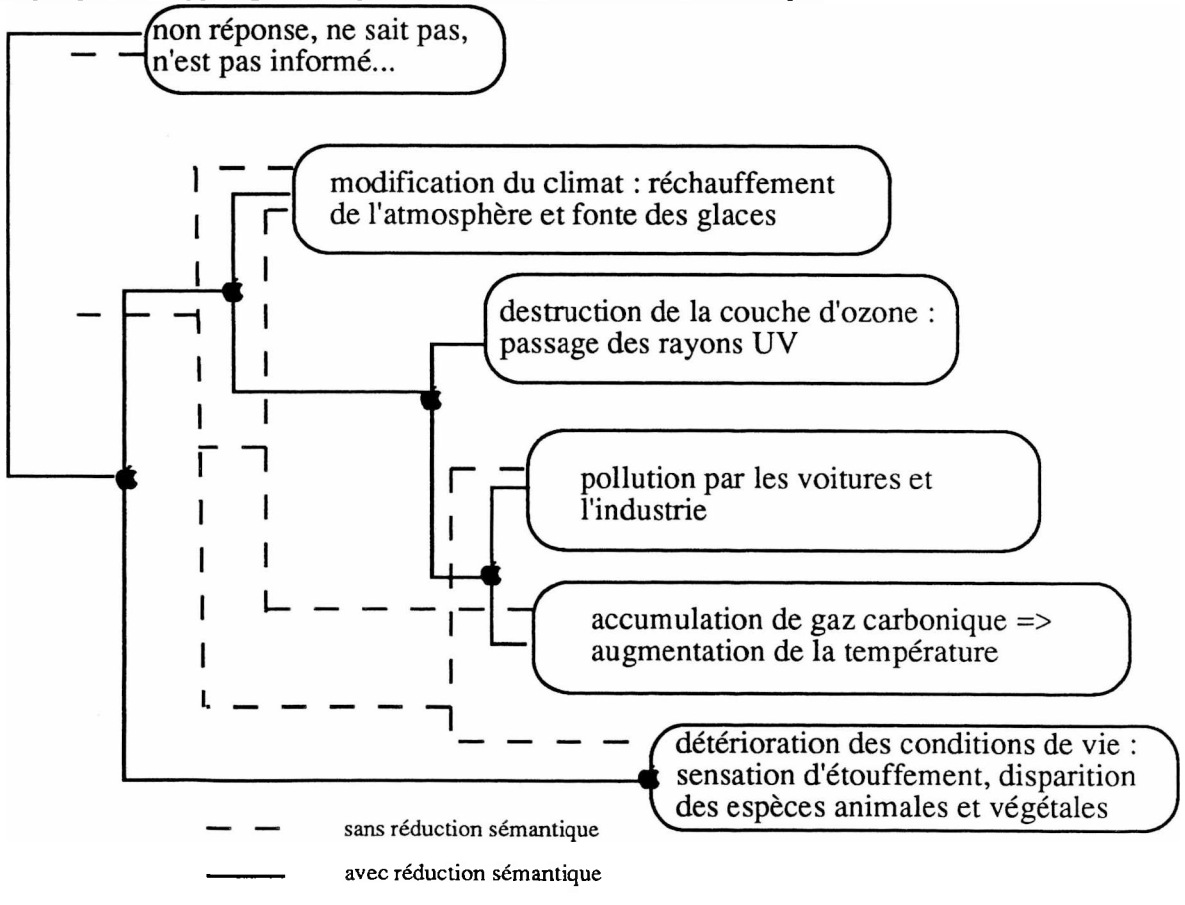
2.3.3. Les effets sur l'analyse lexicale

Les essais que nous avons effectués montrent que, dans l'état actuel, la réduction sémantique change finalement assez peu la nature des classes obtenues, mais qu'elle fait changer l'arbre de classification et la taille des classes. Elle provoque également des glissements sémantiques qui sont difficiles à apprécier : sont-ils favorables ou non à l'analyse ? Il faudra faire des essais sur de nombreux corpus pour se faire une idée précise.
Globalement, les effets de dérive introduits par le réseau sont noyés par les effets de masse (redondance des réponses).

Tableau 6 : Tableau récapitulatif des classifications obtenues selon le mode de réduction

Nom des classes	Lemmatisation ALCESTE	Lemmatisation linguistique analyse syntaxique et lexique	Réduction sémantique
Destruction de la couche d'ozone	15 %	14 %	14 %
Accumulation de gaz carbonique	22 %	22 %	15 %
Modification du climat : réchauffement de l'atmosphère et fonte des glaces	26 %	27 %	20 %
Pollution (voiture et industrie)	12 %	10 %	16 %
Détérioration des conditions de vie (étouffement, disparition des espèces)	16 %	18 %	24 %
Ne sait pas, pas assez informé, aucune idée	4 %	1 %	1 %
U C E non classées	5 %	8 %	

Graphique 11 : Typologie des réponses : effet de la réduction sémantique



Nous venons de voir comment avaient évolué les méthodes pour réduire la taille du vocabulaire. Nous avons surtout constaté que, dans le cas des questions ouvertes, le système d'analyse de cooccurrences est tellement robuste qu'il est finalement assez peu sensible au mode de lemmatisation. Seule la réduction sémantique devrait permettre de sérieuses avancées pour les textes moins redondants que les questions ouvertes, à condition bien sûr de forcer la convergence de la réduction par une prise en compte du contexte beaucoup plus serrée. D'autre part, pour la réduction sémantique, le problème de la polysémie n'a tout simplement pas encore été abordé.

DEUXIÈME PARTIE : DIVERSIFICATION DES CHAMPS D'APPLICATION ET DES OUTILS DE L'ANALYSE LEXICALE

Outre les avancées techniques et linguistiques sur les méthodes de réduction des corpus (au sens large), nous avons eu l'occasion au cours de ces dernières années en faisant des analyses sur des corpus de plus en plus différents (et plus seulement des réponses à des questions ouvertes), en travaillant dans des domaines connexes au langage naturel (c'est-à-dire en sélectionnant des traits pertinents pour l'analyse qui ne sont pas de nature purement lexicale), et en utilisant de nouveaux logiciels, de nous écarter des analyses standard et d'appliquer l'analyse des données textuelles à d'autres problématiques que celles du classement de réponses à des enquêtes d'opinion.

En choisissant ALCESTE, logiciel adapté à l'analyse de textes de longueur et de nature variable et à la sélection de traits pertinents, le champ d'application des techniques d'analyse lexicale s'est considérablement élargi.

En même temps que le champ d'application s'élargissait, il a paru nécessaire d'utiliser d'autres logiciels d'analyse textuelle. En complément du logiciel ALCESTE, on a utilisé de plus en plus fréquemment Hyperbase. Ce logiciel développé en Hypercard par Étienne Brunet[11] est générateur d'hypertextes : le format des textes entrés dans le logiciel permet une circulation libre au sein des textes. Le logiciel est doté d'outils de consultation et de quelques outils statistiques.

Cet élargissement des champs d'application, nous a amené à une réflexion sur :

- la définition des unités d'analyse analysées ;
- les possibilités d'application des méthodes d'analyse lexicale à des descriptions d'objets en langage non naturel.

---

11 BRUNET Etienne, *Hyperbase, Notice d'utilisation*, CUMFID n°17, 84 p.

1. RÉFLEXION SUR LES UNITÉS D'ANALYSE : VERS UNE DÉFINITION RELATIVISTE DU TYPE D'UNITÉ

On peut trouver deux origines au développement de l'analyse des données textuelles comme l'a fait remarquer Max Reinert (1992a) lors d'une conférence à Albi : d'une part la statistique lexicale qui repose sur le comptage de mots ; d'autre part, l'analyse distributionnelle initiée par Bloomfield puis développée par Harris. Au croisement de la statistique et de la linguistique, la statistique lexicale hérite de certaines des problématiques propres à chacune des disciplines, plus particulièrement celle de la définition des unités d'analyse que nous allons examiner maintenant.

Le principe de la statistique est de mettre en évidence des ressemblances et différences entre des individus ou unités statistiques appartenant à une population (Lahlou, Maffre et Beaudouin, 1991). Dans le cas de la statistique textuelle, la population est un ensemble de séquences textuelles que l'on a l'habitude d'appeler corpus. Tout comme pour une population statistique, il faut que le corpus réponde à un critère de cohérence et d'homogénéité qui peut être soit thématique soit formel. Dans le cas des questions ouvertes ou des entretiens semi-directifs, c'est le thème abordé dans la ou les questions qui assure l'homogénéité du corpus ; dans l'étude d'un recueil de sonnets, d'un ensemble de décisions civiles ou d'une rubrique de presse sur une longue période, c'est la cohérence formelle qui justifie la constitution du corpus. Il est donc indispensable avant toute analyse de définir la population, c'est-à-dire le corpus.

La définition des individus ou unités d'analyse est au cœur même de l'analyse statistique et elle est particulièrement importante en statistique textuelle car plus problématique. Dans une enquête classique quantitative, ce sont les individus que l'on compare, c'est-à-dire des "objets" aux frontières relativement bien définies qui, aux yeux de ceux qui les "étudient", ont une cohérence interne. Mais quand on travaille sur des données textuelles, quelle unité statistique doit-on choisir ?
Quand il s'agit de réponses à des questions ouvertes, le problème paraît très simple : l'unité est la réponse de la personne interrogée. Cette unité n'est finalement qu'apparente : une réponse peut être constituée de différents noyaux de sens. Il est alors aussi judicieux de découper les réponses en sous-segments, chacun correspondant à un concept ou à une idée.
Et si l'on étudie un ensemble d'articles, des entretiens semi-directifs, un texte littéraire ou un ensemble de textes littéraires, quelles unités doit-on retenir ? C'est le recours à la linguistique

et aux définitions des unités d'analyse linguistique qui peut nous éclairer dans le choix des unités statistiques.

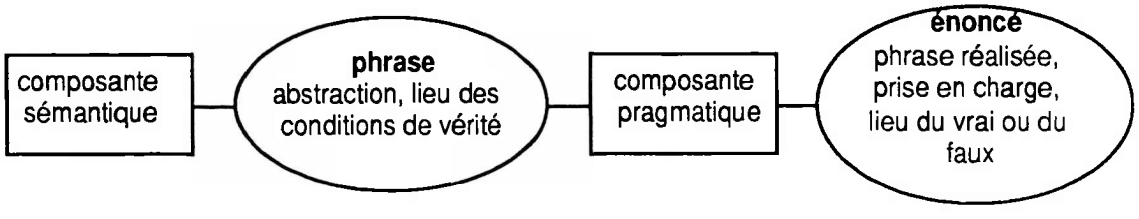
En linguistique, selon le domaine de spécialisation, le type d'unité analysée varie. En lexicologie, on travaille sur le mot, en grammaire sur la phrase, en pragmatique sur l'énoncé, en analyse de la conversation sur les tours de parole, en stylistique sur le texte... La définition de l'unité est en linguistique relative au type de problématique de l'analyste.

Tout comme en linguistique, l'unité statistique textuelle devrait être définie au cas par cas en fonction de ce que l'on cherche à mettre à jour. Malheureusement, le repérage automatique d'unités linguistiques est loin d'être évident et pour le moment, les logiciels existants ne nous offrent pas toutes les possibilités. ALCESTE en caractérisant et en classant des énoncés définit des unités plutôt micro-structurales tandis qu'Hyperbase, en mettant en évidence les spécificités de textes voire de chapitres, travaille plutôt sur des unités macro-structurales. Ce type de découpage correspond à deux manières d'envisager le matériel textuel, ce qui explique la complémentarité des deux logiciels.

1.1. Le découpage alcestien : unité de contexte élémentaire et énoncé

Pour la méthodologie ALCESTE, Max Reinert a choisi comme unité d'analyse l'énoncé. Mais qu'est-ce qu'un énoncé en linguistique ? Et comment peut-on le repérer de manière automatique ?

Pour un certain courant de la linguistique[12], ce qui distingue la phrase de l'énoncé, c'est que la première appartient à la langue tandis que le second est de l'ordre du discours. Clairement, l'énoncé est une phrase réalisée, c'est-à-dire prise en charge par un locuteur en un lieu et temps donné. L'énoncé est au cœur de la pragmatique.



12 Voir en particulier Robert MARTIN, *Pour une logique du sens*, Paris, P. U. F. , 1992.

En signalant que l'énoncé est une phrase réalisée, on sous-entend que l'énoncé est défini par des considérations d'unité syntaxique. L'énoncé est une phrase réalisée, c'est-à-dire une suite de mots situés entre deux points.

Si l'on se place du point de vue du courant de l'énonciation ou de la pragmatique, les définitions sont quelque peu différentes.
Pour Ducrot Oswald (1984), ce ne sont pas des considérations syntaxiques qui définissent l'énoncé, mais des considérations sémantiques et pragmatiques : l'énoncé doit être relativement autonome, c'est-à-dire répondre aux critères de *cohérence* et d'*indépendance.*

> Un interprétant doit, pour segmenter en énoncés un discours donné, admettre que ce découpage reproduit la succession de choix relativement autonomes que le sujet parlant prétend avoir opéré.
>
> (Ducrot, 1984)

Prenons un extrait d'entretien et cherchons à identifier les frontières d'énoncé :

> Et quand Floriane avait 5 ans, je m'étais dit que c'était vraiment la limite à partir de laquelle elle pouvait peut-être accepter que je m'en aille, quoiqu'elle ne me voie pas tout le temps, qu'elle me voie une semaine sur deux. // Il est à remarquer qu'on a été tout à fait correct l'un et l'autre vis-à-vis de Floriane. On ne l'a pas utilisée comme une espèce de balle qu'on pourrait se relancer. Elle n'a pas été prise en otage du tout // mais moi j'ai très très mal vécu de l'avoir une semaine sur deux et de ne pas l'avoir tout le temps, bien que cela apporte des compensations.
>
> Extrait d'entretien du CES sur la prévention-sida

Le signe "//" signale une frontière entre deux énoncés, qui ne coïncide pas toujours avec les fins de phrases. Ainsi, l'énoncé n'est pas forcément réductible à la phrase ou à toute autre unité langagière comme on peut le voir sur l'extrait d'entretien. Seul le recours au sens permet donc d'en définir les limites. L'énoncé porte la marque du sujet qui le prend en charge et qui dit quelque chose sur le monde. Pour Max Reinert, dans la lignée de tous les philosophes réalistes qui considèrent que la parole renvoie à la réalité du monde, c'est dans l'unité représentée par l'énoncé que s'exprime le mieux une "vision du monde".

Mais en retenant une telle définition, il devient difficile de découper de manière automatique un texte en énoncés. En effet, comment définir des critères qui, analysant la cohérence pragmatique d'un énoncé, peuvent en fixer les frontières ?

Nous avons vu que l'énoncé ne correspond pas à la phrase (on comparera par exemple la différence entre l'énoncé synthétique d'une réponse à une question ouverte et la longueur d'une réponse lors d'un entretien semi-directif où le même contenu conceptuel est exprimé

avec beaucoup plus de mots, d'exemples illustratifs...). Il fallait donc inventer un système automatique pour le découpage des énoncés. Le logiciel ALCESTE découpe le texte en unités de contexte composées d'un certain nombre de lignes. Ces unités respectent plus ou moins la ponctuation. Comme ce découpage est assurément arbitraire, et qu'il ne peut y avoir de coïncidence entre un découpage d'unités de longueur constantes et des énoncés qui sont de longueur variable, le logiciel propose de comparer deux types de découpage en faisant varier la longueur des séquences. Tous les calculs d'analyse des données s'appuient sur les deux découpages et, en fin d'analyse, la comparaison des résultats permet de conserver les noyaux communs qui sont l'expression de certaines visions du monde.

Ce choix de découpage statistique a en outre deux avantages :
- il permet de comparer des textes de longueur variable puisque chaque texte est découpé en fragments de longueur similaire (un certain nombre de lignes) ;
- chaque unité de contexte ou pseudo-énoncé peut-être caractérisé par un ensemble d'indicateurs (appelées variables étoilées dans ALCESTE). Pour des questions ouvertes, ce sont les caractéristiques socio-démographiques, comportementales... de la personne interrogée et, pour des textes littéraires, l'auteur, la date de parution, le titre de l'ouvrage, du chapitre auquel appartient l'unité de contexte.

Par exemple, le corpus sur les Caisses d'Épargne était constitué de soixante-dix extraits de la *Revue de l'année,* rubrique du Journal des Caisses d'épargne qui faisait le bilan de l'activité de l'année écoulée. Chaque unité d'analyse était indicée par l'année de parution de l'article, ce qui permettait en fin d'analyse de voir à quelle période correspondait chaque type de discours (Annexe 3).
Dans l'étude sur les outils de communication nutritionnelle faite pour Agnès Mignonac Conseil, chaque document était considéré comme une unité d'analyse et identifié par le type d'émetteur (entreprise, médecin, diététicien...), le type de récepteur, la date de diffusion... ce qui permettait de mieux caractériser les types de discours sur la diététique et de voir leur évolution (Annexe 2).

1.2. Le découpage Hyperbase : unités "naturelles" de découpage

Dans le logiciel Hyperbase, principalement conçu pour l'étude des textes littéraires, l'unité retenue est l'ouvrage, éventuellement ses sections.
Pour François Rastier, la seule unité valable comme base de l'analyse et en particulier pour l'étude du sens est le texte et non pas, comme le font habituellement grammairiens et linguistes, la phrase. Or si l'on veut retenir ce type d'unités, il faut pour cela disposer de quantités de textes importantes et faire en sorte que tous les calculs tiennent compte des longueurs variables des textes.
C'est le principe même du logiciel d'Étienne Brunet de permettre des comparaisons entre les unités naturelles d'un recueil (chapitre, section, parties...) et même de comparer une oeuvre ou plusieurs œuvres à un ensemble d'oeuvres. Hyperbase contient en mémoire la fréquence des termes sur la plupart des oeuvres littéraires de 1823 à 1923 (corpus construit pour servir de base d'exemples au Trésor de la Langue Française (TLF)). Ainsi pour étudier les particularités du recueil *Gravitations* de Jules Supervielle, ont été mis en évidence les termes qui étaient sur et sous-représentés dans le recueil par rapport au TLF et qui définissent les particularités d'une certaine vision du monde. On a aussi pu mettre en évidence les spécificités lexicales de chacune des sous-parties du recueil (une sous partie est comparée à l'ensemble des autres (Beaudouin, 1993)).

C'est une approche macrostructurale que propose Hyperbase dans ses analyses statistiques puisque l'on compare toujours des unités textuelles et "naturelles". Mais l'intérêt d'Hyperbase ne repose pas uniquement sur ses outils statistiques. Les outils de consultation apportent souvent un éclairage enrichissant aux résultats d'ALCESTE. On peut en particulier retrouver tous les contextes d'utilisation de certains termes.
Ainsi pour les Caisses d'épargne, on a pu voir que le terme *institution* fréquemment utilisé pour désigner les Caisses d'Épargne jusque dans l'entre-deux guerres, disparaît du vocabulaire du *Journal* pour réapparaître encore plus fréquemment à partir des années 60, alors même que l'évolution du contexte économique obligeait les Caisses à renoncer à cette identité d'Institution et à se considérer comme une entreprise prise au jeu de la concurrence (Annexe 3).

En résumé, la complémentarité d'ALCESTE et d'Hyperbase vient de la différence dans la définition des unités d'analyse : point de vue micro-structural (s'appuyant sur les énoncés) opposé à un point de vue macrostructural (découpage en unités naturelles).

2. L'APPLICATION GÉNÉRALE DE LA MÉTHODE À DES DESCRIPTIONS D'OBJETS EN LANGAGE NON NATUREL

2.1. Le principe de la description par traits

La description des objets se fait forcément par une combinatoire de traits qui sont les éléments de langage de description utilisé. Le cas particulier de la description par des fichiers issus de systèmes de description systématiques (typiquement, les questionnaires fermés) produit des descriptions de populations sous la forme de matrices pleines, où les lignes sont les individus et les colonnes les traits. Les méthodes usuelles de la statistique descriptive et multivariée sont bien adaptées à ces descriptions saturées, dans lesquelles chaque trait est décrit pour chaque objet.
En revanche, les descriptions naturelles se font en général dans un espace de variables non délimité a priori, ou du moins dans lequel les objets ne sont décrits que par une petite partie des traits possibles, ceux qui sont pertinents ou accessibles au système de description. Quand on essaie de comparer des objets décrits de cette façon *naturelle*, sous forme de matrices individus-traits, on obtient des tableaux hypercreux, surtout constitués de zéros, car chaque individu est non-décrit sur la plupart des traits. Par ailleurs, comme il n'existe pas alors de protocole de description standardisé, l'ordre d'occurrence des différents traits dans une description est variable. Ce type de tableau est extrêmement difficile à traiter avec les méthodes usuelles, car les absences de descriptions prennent le statut de *valeur manquante* (par rapport à une structure idéale de tableau plein). Dans l'analyse, les logiciels ont alors tendance à éliminer les individus qui ont trop de valeurs manquantes, et donc, en l'espèce, la plupart des individus.

Or, ce cas est précisément celui des descriptions en langue naturelle, où chaque objet n'est décrit qu'avec une petite partie du vocabulaire. Les outils qui ont été mis en place pour l'analyse lexicale peuvent donc servir à analyser les matrices hypercreuses d'une manière générale. Nous avons essayé de les appliquer à trois cas particuliers rencontrés dans notre pratique, et nous nous en sommes trouvés fort satisfaits. Par ailleurs, ceci nous a donné l'idée d'appliquer ces méthodes aux fichiers fusionnés, qui sont une forme particulière de matrice hypercreuse : on peut considérer un fichier fusionné comme une matrice "creuse par blocs".

2.2. Le classement des objets décrits par des traits "en désordre"

Voyons d'abord les trois applications de la méthode lexicale à des fichiers issus de données quantitatives. Le principe de l'analyse globale qui a été suivi dans le premier cas consiste à faire des classifications "en cascade" (Lahlou, 1988, annexes techniques) qui permettent un meilleur contrôle de l'interprétation des classes obtenues par analyse factorielle. Des individus sont interrogés sur divers aspects d'un problème donné. Les réponses portent donc sur un grand nombre de variables.
L'approche classique consiste à prendre toutes les variables pertinentes et à les introduire comme variables actives dans une analyse unique, pour faire une classification dont on interprète ensuite les classes. L'inconvénient de cette approche est qu'elle contribue à augmenter le bruit (c'est-à-dire les corrélations non interprétables) en introduisant des corrélations entre *toutes* les variables, et en les traitant sur le même plan dans l'analyse.

La classification en cascade cherche à limiter cette difficulté. Elle consiste à :

- séparer les variables en sous-domaines séparés, qui correspondent intuitivement chacun à un aspect particulier du problème (problématique locale) ;
- opérer indépendamment une classification sur chacun de ces sous-domaines, où l'interprétation est relativement facile ;
- à identifier chaque individu par sa position dans chacune de ces classifications locales, qui deviennent alors des variables synthétiques ;
- à opérer une classification globale sur la base de ces nouvelles variables synthétiques.

L'expérience montre que la stabilité et l'interprétation des classes finales obtenues sont bien meilleures que par la méthode brutale, car on a, de fait, séparé dans l'analyse le bruit résultant de la variance interne à chaque problématique locale (qui est assimilable à une variance intra-classe) et on l'a éliminée de l'analyse finale.

Dans le cas du petit déjeuner, nous avions interrogé en ouvert les individus, notamment sur les cinq aspects locaux suivants :

```
* un petit déjeuner idéal, à quoi ça vous fait penser?
* décrivez votre petit déjeuner habituel
        - en semaine
              ambiance
              produits consommés
        - en week-end, en vacances
              ambiance (6)
              produits consommés (9).
```

Nous avons opéré une analyse sur chacune de ces problématiques locales, et nous voulions une vision globale.

A l'issue des cinq analyses, nous avions obtenu cinq classements des individus. Afin d'exploiter conjointement ces résultats, nous avons effectué, à l'aide du progiciel SAS, une ACM sur le numéro des classes des individus. Nous avons donc créé un fichier contenant 6 variables : l'identifiant de l'individu et, pour chacune de ses cinq réponses, le nom de la classe dans laquelle celle-ci avait été rangé. Il s'est avéré que cette analyse n'était pas interprétable, les axes représentaient tous un pourcentage d'inertie du nuage voisin de 4%. Ceci peut être expliqué par la présence de certaines classes parasites (de faible effectif) qui pouvaient créer un axe à elles seules.

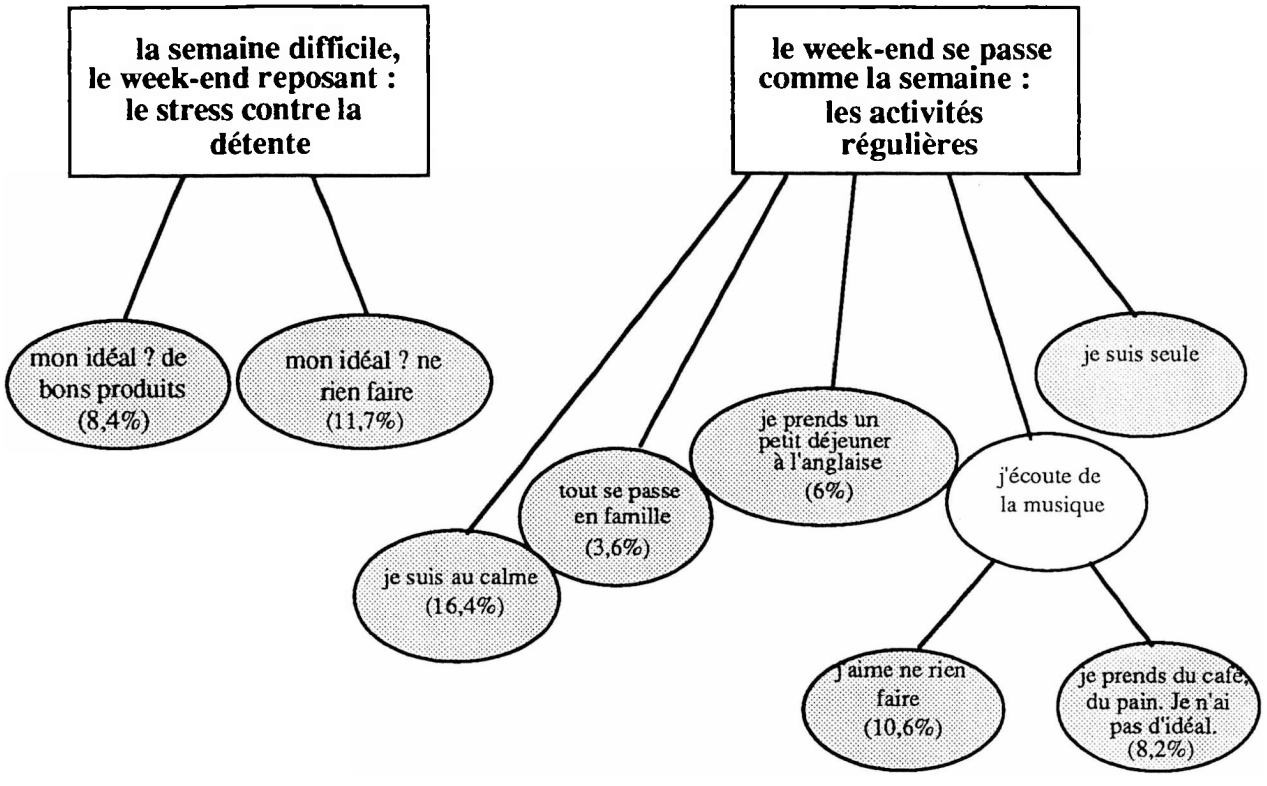
Pour remédier à ce problème, une première option était de reventiler au hasard les classes de trop faible effectif (en fixant le seuil à 2% du total des individus par exemple) comme le fait le logiciel SPAD. Nous avons choisi une autre optique, à savoir effectuer au préalable sur nos classes une classification descendante hiérarchique. En effet celle-ci a le mérite de séparer à chaque étape le nuage en deux parties opposées. Ainsi ces classes de réponses parasites seraient séparées dès le début du reste du nuage. C'est la méthode adoptée par le logiciel ALCESTE.

Utiliser pour cette étude ALCESTE résolvait également un autre type de problème, celui de la lisibilité des résultats. En effet pour interpréter la création des axes, il fallait sans arrêt se reporter au texte et aux résultats de l'analyse textuelle. En utilisant ALCESTE, nous donnions un intitulé aux classes, par exemple ambiance_détendue, la lecture était alors immédiate. Nous avons choisi de regrouper les classes de faible effectif et les classes non directement interprétables sous le même intitulé : résidu.

Chacun des cinq classements comporte deux modalités supplémentaires : _zéro, qui correspond aux individus non classés et _blanc qui correspond à quelques erreurs de fichier, c'est-à-dire à des réponses que l'on n'a pas pu retrouver. (Les non-réponses sont comptabilisées dans la catégorie non classées par ALCESTE). Nous avons également introduit les lieux de la semaine et du week-end en variables supplémentaires.

Les résultats sont dans l'ensemble très satisfaisants. L'arbre est régulier, et seules deux classes sont d'effectif très faible (respectivement 15 et 12 individus). On trouve deux classes d'individus que l'on n'a pas réussi à caractériser précisément. Elles correspondent à des individus indécis ou inclassables (c'est-à-dire non classés ou situés dans la classe artificielle résidu pour plusieurs réponses). Elles comportent respectivement 69 et 117 individus. Seuls 23 individus ne sont pas classés. Nous pouvons donc interpréter le classement de 770 individus soit près de 77% de l'échantillon de départ.

Graphique 12: Typologie sur l'ensemble du questionnaire



Le résultat le plus marquant de cette analyse est que ce sont surtout les ambiances respectives de la semaine et du week-end (leur accord ou leur opposition) qui ont créé les classes, l'idéal, tout comme les produits consommés ne jouant qu'un faible rôle.

On observe très nettement deux types de regroupements. Celui qui caractérise une opposition d'ambiance entre la semaine et le week-end, celui qui caractérise un accord d'ambiance entre la semaine et le week-end. Notons que l'idéal est proche du petit déjeuner dominical dans le cas d'une opposition entre la semaine et le week-end.

(Beaudouin et al., 1993)

L'application des méthodes d'analyse lexicale pour traiter des fichiers fermés comme s'il s'agissait de réponses en langue libre s'est donc montrée tout à fait rentable ici. Le même résultat aurait pu être obtenu avec les méthodes traditionnelles, mais aurait nécessité des opérations de manutention de fichier et de recodage considérables et fastidieuses, et n'aurait pas éliminé les artefacts dus aux valeurs manquantes.

2.3. Le classement d'objets décrits par des séquences de traits

Dans les méthodes d'analyse des données textuelles, chaque énoncé ou unité textuelle est considéré comme une combinatoire de traits. L'analyse porte sur une matrice qui croise les unités textuelles et les mots employés, matrice hypercreuse qui nécessite des types de traitements particuliers.
Mais dans l'analyse de certains corpus, l'expérience nous a peu à peu montré qu'il était indispensable de sélectionner certains traits (ou mots) pertinents pour aboutir à une description plus spécifique de l'objet. Si pour l'analyse d'une question ouverte ou d'un texte littéraire, chaque mot a une importance dans la mesure où, dans la phase d'interprétation, chacun intervient dans la reconstruction du sens, dans le cas de certaines analyses portant plus sur la structure formelle ou énonciative que sur le contenu, il devient indispensable de repérer ces traits pertinents qui constituent la charpente de l'énoncé et de faire porter les analyses statistiques exclusivement sur eux. Le corpus initial est alors projeté dans un langage constitué uniquement de ces traits pertinents. Nous avons eu l'occasion avec le CERCRID de mettre au point des analyses de ce type.

Dans le cadre d'une étude sur le langage du dispositif des décisions judiciaires civiles[13] réalisée par Évelyne Serverin du CERCRID et Sylvie Bruxelles du CNRS, nous avons cherché à constituer une typologie des décisions judiciaires civiles. Cette étude avait pour but de venir en soutien à un groupe de recherche chargé d'élaborer une nouvelle nomenclature des décisions judiciaires civiles.
Pour élaborer ce type de nomenclature, il était important de repérer la structure des décisions des différentes juridictions et ceci, comme pour toute analyse de structure, indépendamment de l'affaire traitée. Pour identifier les points communs à différentes décisions au niveau lexical, il était indispensable de repérer les éléments constitutifs de la structure.

> La "nomenclature des dispositifs" associe deux catégories d'informations :
> - *la nature de l'acte* qui dessaisit le juge de l'affaire, ou en suspend le cours,
> - *l'effet juridique* de la décision, défini comme la situation créée par le jugement.
>
> (Serverin, Bruxelles, 1992)

Dans cette optique, il était particulièrement intéressant d'analyser la décision en tant qu'acte de langage. Or la catégorie syntaxique la plus apte à exprimer un acte est le verbe et plus particulièrement le verbe performatif, c'est-à-dire le verbe qui constitue simultanément l'acte auquel il se réfère.

---

[13] SERVERIN Evelyne, BRUXELLES, Sylvie.- *Le langage du dispositif des décisions judiciaires civiles: Contribution à l'étude de l'acte décisionnel*, CERCRID-CNRS, août 1992.

> **Verbe** :*I. (1170; déjà en lat.). Mot qui exprime un **dynamisme (action, état, devenir)** et qui présente dans de nombreuses langues - notamment en français - un système complexe de formes (- Conjugaison).
>
> Le Robert Électronique

D'où l'idée de caractériser les décisions par les types de verbes employés.
L'analyse a porté sur un mois de décisions rendues par quatre juridictions représentant chacune une des grandes catégories de tribunaux statuant au fond (Cour d'appel, Tribunal d'Instance, Tribunal de Grande Instance, Conseil des Prud'hommes). L'unité d'analyse était la décision civile proprement dite caractérisée par la juridiction dont elle émanait et par les verbes qui y étaient employés.

Nous avons vu qu'en statistique usuelle un objet est décrit par une combinaison (et non par un arrangement ordonné) de traits. En appliquant les méthodes statistiques au texte, on perd les informations sur la séquentialité (ordre des mots, structure syntaxique...). Or, il était indispensable dans cette étude de pouvoir reconstituer l'ordre d'apparition des verbes. Une décision ne pouvait être caractérisée par un groupement non ordonné de traits, mais bel et bien par une séquence de verbes.

Concrètement, dans l'analyse, on n'a conservé comme variables actives que les verbes ou formes verbales apparaissant dans les dispositifs. Au lieu de considérer les verbes pris isolément comme variables, on a retenu les couples de verbes, c'est-à-dire des suites de deux verbes .
En application de ce principe, la décision suivante de la Cour d'appel de la chambre de la famille :

> "Par ces motifs, confirme l'ordonnance déférée en toutes ces dispositions, et ce compris les dépens ; déboute (X) de sa demande formée sur les dispositions de l'article 700 du nouveau code de procédure civile ; condamne (Y) aux dépens d'appel ; autorise l'avoué adverse à recouvrer contre la partie condamnée ceux des dépens avancés sans avoir reçu de provision conformément aux dispositions de l'article 699 du nouveau code de procédure civile."

sera caractérisée par les couples de verbes suivants :

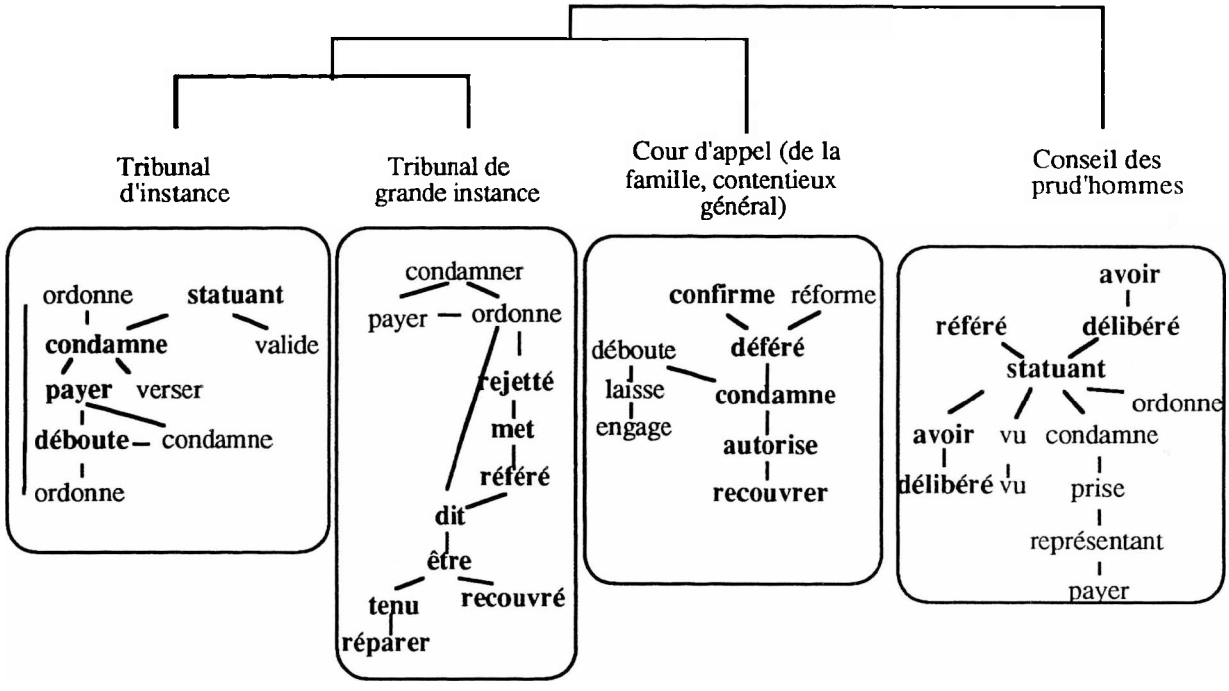
> confirme-déférée déférée-déboute déboute-formée (formée-condamne) condamne-autorise autorise-recouvrer recouvrer-condamnée (condamnée-avancés) (avancés-avoir) avoir-reçu

Les couples qui ont une fréquence trop faible dans l'ensemble du corpus ne sont pas retenus : couples entre parenthèses.

Chaque décision est donc caractérisée par l'ensemble des couples qui y apparaissent. Pour constituer une typologie de ces décisions, un tableau à double entrée est établi, qui contient en ligne les décisions et en colonne les couples de verbes. Sur la base de ce tableau, on a effectué une classification descendante hiérarchique avec ALCESTE.

Ce qui est remarquable dans cette analyse c'est que la typologie obtenue recouvre à plus de 90 % les différentes catégories de tribunaux. En effet, on obtient quatre classes de décisions (caractérisées par des séquences de verbes) et chacune d'entre elles regroupe les décisions d'un type de tribunal.

Graphique 13 : Typologie des décisions judiciaires civiles



Les séquences de verbes sont les marques de l'écriture de chacune des juridictions.

> Les décisions du **Conseil des Prud'hommes** se distribuent autour de la mention qui introduit de courtes séquences pour les formes "vu", "ordonne" et une séquence plus longue en cas de condamnation.
> Les **Cours d'Appel** placent en tête, conformément à leur position de juridiction de recours, les verbes qui indiquent le sort réservé à la décision dont la position de décision contestée est marquée par l'emploi du participe "déféré".
> Le **Tribunal de Grande Instance** centre la séquence de ses verbes sur l'énoncé du principe de l'obligation à réparation, suivi d'une condamnation à payer.
> Enfin, le **Tribunal d'Instance** place la condamnation au paiement au centre de la décision, engendrant d'autres séquences plus ou moins longues de condamnation.
> L'écriture du dispositif semble donc varier d'abord dans les enchaînements de formes, puis dans l'emploi des verbes, sur fond d'une permanence de séquences relatives à la condamnation.
>
> (Serverin, Bruxelles, 1992)

A travers cette analyse, on a montré que la sélection de traits pertinents permettait de décrire les objets d'un point de vue formel et énonciatif. Mais aussi qu'il était possible, avec les outils d'analyse lexicale, de considérer des séquences de traits pertinents pour le problème posé plutôt qu'une combinatoire globale pour décrire un objet. Ce "recodage ouvert" est à envisager dans tous les cas où le corpus initial n'a pas été recueilli *spécifiquement* pour l'étude en cours.
L'analyste ne doit pas se laisser imposer le choix du langage de description par les seules conditions matérielles de recueil d'information. Le matériau de base n'est pas sacré et intangible, c'est une matière brute que l'on peut légitimement traiter de mille façons pourvu qu'elles se justifient par la problématique.

2.4. Le classement des objets complexes en transformant des descriptions codées par la concaténation des libellés des codes

Une autre application a consisté à classer des produits alimentaires décrits de façon structurée, mais dans un langage très vaste : EUROLANGUAL (pour une description détaillée de ce langage : (McCann et al., 1988 ; CFSAN, 1990 a et b ; Feinberg et al, 1991 ; Lahlou, Maffre, Beaudouin, 1991 ; Lahlou, octobre 1992 ; CREDOC, 1993 ; Litman, 1993).

Un problème, quasiment impossible à gérer en fichier rectangulaire, résulte du fait que la description de ces objets est une description "à tiroirs" : chaque aliment est considéré non seulement comme un objet en tant que tel, mais aussi comme une recette, c'est-à-dire comme une combinaison d'ingrédients.

Nous avons tourné la difficulté en transformant la description structurée sous forme de codes en une description en langage libre, en remplaçant les codes par leurs libellés.

Ainsi, par exemple le '*potage LIEBIG tomates*', initialement codé de la façon suivante (en fait, seuls les codes A0198, B1276 etc. et les identificateurs numériques des ingrédients de la recette constituent le codage figurant dans le fichier) :

PRODUIT : 33981 **GROUPE** : BSN **SOCIÉTÉ** : SEGMA-LIEBIG-MAILLE

Libellé : LIEBIG POTAGES TOMATES 30 CL BRIK

Libellé complémentaire : LIEBIG PUR SOUP TOMATES BRIK

FACETTES EUROLANGUAL

Code Libellé
A0198 SOUPE
B1276 TOMATE (LYCOPERSICUM ESCULENTUM)
C0139 FRUIT NON PELÉ SANS TROGNON NI NOYAU NI GRAINE
F0014 TRANSFORMATION THERMIQUE COMPLETE
J0120 CONSERVE PAR TRAITEMENT THERMIQUE
K0027 CONDITIONNE SOUS VIDE
M0153 RÉCIPIENT EN CARTON PLASTIFIE
I861 POTAGES PRÉPARÉS
Y6310 POTAGES EN BRIQUES
E0130 LIQUIDE
G0013 CUIT DANS L'EAU OU UN LIQUIDE AQUEUX
H0158 SACCHAROSE AJOUTE
H0163 ENRICHI EN VITAMINES
H0173 SALE
H0227 AROME OU ÉPICE OU HERBE AROMATIQUE AJOUTE
H0263 CORPS GRAS VÉGÉTAL AJOUTE
H0271 BEURRE AJOUTE
H0279 AMIDON MODIFIE AJOUTE
H0280 SIROP DE SUCRE OU SOLIDES DU SIROP AJOUTE
H0349 OIGNON AJOUTE
N0017 POLYÉTHYLÈNE
P0024 ALIMENT SANS PRÉCISION D'AGE
P0095 MENTION RELATIVE AUX VITAMINES OU AUX MINÉRAUX

INGRÉDIENTS DE PREMIER NIVEAU
BEURRE
SUCRE
SEL
EXTRAITS AROMATIQUES VÉGÉTAUX
SIROP GLUCOSE
AMIDON MODIFIE
ÉPICES ET PLANTES AROMATIQUES
TOMATES
HUILE VÉGÉTALE
VITAMINE C
VITAMINE PP
VITAMINE B5
VITAMINE B1
VITAMINE A
EAU
OIGNONS PELÉS

devenait :

> LIEBIG potages tomates un litre brik, tomate, conservé par traitement thermique, conditionné sous vide, aliment sans précision d'âge, liquide, polyéthylène, beurre, sucre, sel, farine de blé, amidon modifié, épices et plantes aromatiques, tomates, huile végétale, lait écrémé, préparation aromatique, eau, oignons pelés

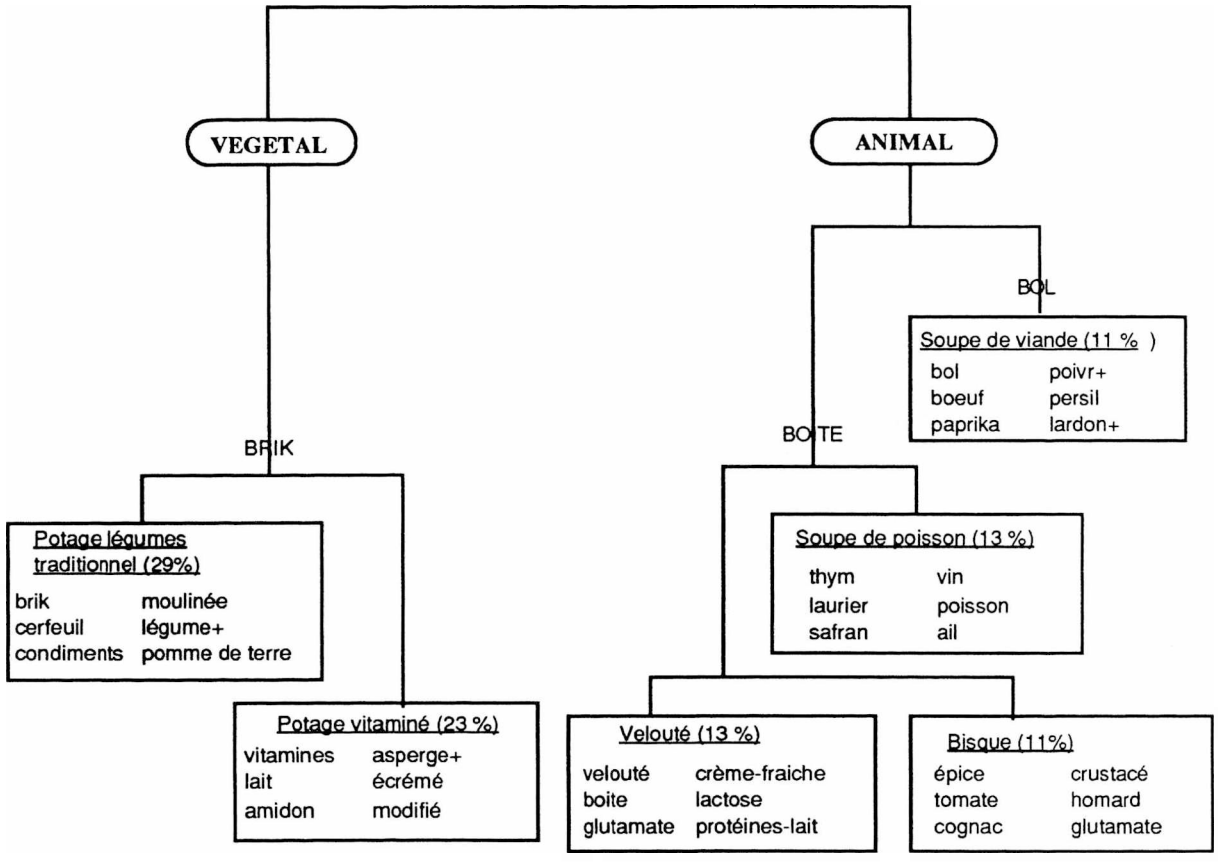
Le corpus de mots que nous avons analysé est donc constitué par les libellés des potages, auxquels sont accolés les libellés de certains codes LANGUAL affectés aux produits ainsi que la liste des ingrédients de ces produits.

Le résultat est une classification des aliments d'un point de vue technologique, à partir de leur description technique[14].

[14] Cette classification a été réalisée par Pascale Hébel, du département Prospective de la Consommation du CREDOC.

Nous avons fait une typologie des différents produits en classant ensemble ceux qui ont une description (ingrédients, libellés ou codes LANGUAL) proche. Seront regroupés les potages qui ont un libellé descriptif contenant les mêmes mots et donc par extension les mêmes ingrédients ou les mêmes codes LANGUAL.

Graphique 14 : Typologie des potages et soupes liquides



La méthode à base d'analyse lexicale met en évidence deux pôles et six classes.

Le premier pôle regroupe les potages dont les ingrédients principaux sont des légumes. Le conditionnement principal est la brik. Ce pôle se divise en deux classes :

- les potages traditionnels contenant très peu d'additifs ;
- les potages à teneur garantie en vitamines.

Le deuxième pôle regroupe globalement les potages à base de denrées animales. La séparation principale se fait par le type de conditionnement et aussi par une distinction entre denrées carnées et produits de la mer. Quatre classes font partie de ce pôle :

- les soupes de viande conditionnées dans des bols ;
- les soupes de poisson additionnées d'épices et d'aromates ;
- les bisques de crustacés ;
- les veloutés : ce sont des potages, en réalité, à base de légumes (d'après les libellés et les recettes) qui contiennent de la crème. Nous expliquons leur présence dans le pôle animal surtout par le fait qu'ils sont tous conditionnés en boîte : cette caractéristique a été déterminante.

(Annexe 4).

On trouvera, en Annexe 4, un descriptif des six classes de la typologie, caractérisées par :

- leur vocabulaire caractéristique constitué des mots caractéristiques lemmatisés c'est-à-dire ramenés à une même racine (par exemple : *légume+* désigne les mots *légume* et *légumes*) ;
- les marques les plus représentatives de la classe ;
- les produits caractéristiques.

Voici, à titre d'exemple, une des classes :

Classe 1 : Potage de légumes traditionnel (29 %)

Ces potages sont élaborés à partir de légumes que l'on peut qualifier de "classiques" (pommes de terre, épinard, légumes variés...). Ils correspondent aux soupes que les consommateurs mangent chez eux. Ils contiennent peu d'additifs comme le proclame le slogan KNORR "Tous les ingrédients d'une soupe faite à la maison et rien d'autre ".

Mots caractéristiques
brik, cerfeuil, condiments, moulinée, légume, pomme de terre, brocoli, beurre, variés, épinard, cresson.

Marques caractéristiques
MAGGI, DIET SOUP, KNORR.

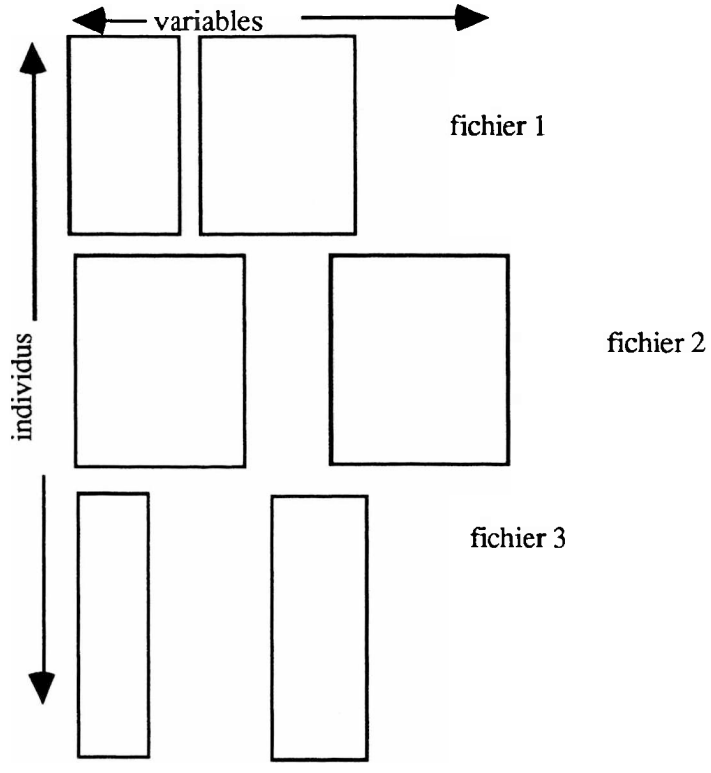
Produits caractéristiques
* MAGGI potages poireaux pommes de terre un litre brik, poireau, conservé par traitement thermique, conditionné sous vide, aliment sans précision d'âge, liquide très visqueux sans particule visible, polyéthylène, beurre, sel, poireaux, eau, poivre, pommes de terre pelées.
* MAGGI potages cresson un litre brik, cresson des jardins, conservé par traitement thermique, conditionné sous vide, aliment sans précision d'âge, liquide peu visqueux, polyéthylène, sel, crème, eau, lait, cresson, pommes de terre pelées.
* KNORR potages champignons un litre brik, champignon, conservé par traitement thermique, conditionné sous vide, aliment sans précision d'âge, liquide, polyéthylène, beurre, persil, sel, crème, champignons, eau, poivre, pommes de terre pelées, ail pelé
* DIET SOUP potages poireaux pommes de terre un litre brik, poireau, conservé par traitement thermique, conditionné sous vide, aliment sans précision d'âge, liquide très visqueux sans particule visible, polyéthylène, beurre, sel, poireaux, eau, poivre, pommes de terre pelées

(Annexe 4).

2.5. Le travail sur les fichiers fusionnés

Les fichiers fusionnés résultent de la concaténation de plusieurs fichiers différents qui possèdent des variables communes.

Ces fichiers peuvent être schématisés de la manière suivante :



Un tel fichier peut être considéré comme hypercreux par blocs.
Les blocs représentent ici des groupes de codes individus x variables. Le cas le plus courant est celui où les fichiers fusionnés ont en commun des descripteurs socio-démographiques d'individus ("caractères communs"), tandis que la plupart des variables varient. La procédure habituelle lors des fusions consiste à attribuer à un individu, là où il a des valeurs manquantes, la valeur que prend cette variable pour un autre individu identique ou proche sur les "caractères communs". Cette méthode possède de nombreux inconvénients bien connus sur lesquels nous ne reviendrons pas et qu'il est difficile de pallier.

Il existe par ailleurs des traits qui sont identiques, mais qui ne sont pas reconnus comme tels lors de la fusion. En effet, ces traits apparaissent dans les fichiers sous la forme de modalités des variables. Par exemple, le trait "marié" peut être une partie d'une modalité dans une variable fermée constituée de :

```
État matrimonial :
- célibataire (E1)
- marié ou concubin (E2)
- veuf ou divorcé (E32)
```

et apparaître dans un autre fichier sous la modalité E'2 d'une variable :

```
État matrimonial ' :
- célibataire (E'1)
- marié (E'2)
- vivant en concubinage (E'3)
- veuf (E'4)
- divorcé (E'5)
```

Un cas plus pervers consiste en un trait qui apparaît de manière à peu près identique dans deux enquêtes en compagnie de traits différents. Par exemple :

```
Votez-vous plutôt :
- pour des partis de gauche
- pour des partis de droite
- autre

Quelles sont vos sympathies politiques ? (réponse ouverte)
```

La post-codification et la remise en forme de tels fichiers posent d'énormes problèmes.

Nous proposons de considérer chaque individu comme une description "en clair" des résultats de l'enquête, en remplaçant chaque modalité par son libellé en langue naturelle. Le fichier total peut alors être traité par analyse lexicale. On récupérera ainsi, dans l'exemple ci-dessus, le libellé "marié" dans les deux cas.

Cette approche est, naturellement, illicite sur le strict plan méthodologique : on sait bien que le questionnement induit des biais dans les réponses, et que des questions différentes ne sont pas strictement comparables. Néanmoins, poussé à l'extrême, ce raisonnement invaliderait toute approche statistique, puisque, pour une même enquête, chaque enquêté ayant été questionné dans un contexte particulier et par un enquêteur différent, on ne peut pas strictement considérer que la question est la même. La méthode que nous proposons est donc simplement une approximation supplémentaire. Il ne faudrait en user qu'avec d'extrêmes précautions, et à des fins exploratoires, quitte à vérifier ensuite proprement les idées qui peuvent émerger de ce premier traitement. Néanmoins, un peu d'information vaut parfois mieux que pas d'information du tout et, sur le plan technique, cette procédure est infiniment plus économique qu'un recodage, elle peut être effectuée avec les procédures de "change all" sur les fichiers bruts. On peut d'ailleurs l'améliorer considérablement si l'on dispose de réseaux sémantiques, .

Le dernier avantage de cette procédure est la possibilité, jusqu'ici techniquement très ardue à réaliser, de se reporter facilement aux descriptifs complets des individus d'origine lors des dépouillements d'analyse, puisque, dans les classifications, chaque individu sera décrit non

seulement par son identifiant, comme c'est le cas habituellement, mais par une description *immédiatement lisible* de toutes ses réponses ; puisque chaque individu est un texte.

Nous n'avons pas encore eu l'occasion d'expérimenter cette approche, mais elle semble idéale pour traiter des fichiers d'enquêtes répétitives sur longue période, où les questions varient un peu dans leur libellé et leurs variables, cas qui est celui de tous les systèmes statistiques publics lourds (enquêtes INSEE, Enquête Aspirations et Conditions de Vie des Français du CREDOC). Par ailleurs, les procédures de transfert entre notre chaîne d'analyse lexicale et le logiciel SAS mises au point dans notre département autorisent un traitement en parallèle des données quantitatives et lexicales (en transférant les mots sous la forme de variables colonnes en SAS). Ceci permet, une fois terminée la phase exploratoire sur le fichier "lexicalisé", de réaliser les traitements de vérification avec les méthodes statistiques traditionnelles.

TROISIÈME PARTIE : L'INTERPRÉTATION EN ANALYSE LEXICALE

La philosophie du traitement des données par l'analyse des données à la française est plutôt euclidienne. Elle consiste à réaliser des opérations de classement ou de réarrangement géométrique sur des "nuages de points", représentant les observations dans un espace euclidien.

La philosophie qui préside au traitement lexical tel que nous le pratiquons est, elle, d'esprit plus combinatoire. On peut en effet considérer chaque morceau du corpus (individu statistique) comme une *combinaison* de traits (les mots) tirés avec remise dans un thesaurus global (le vocabulaire du corpus). L'observation des séquences obtenues, qu'un examen même superficiel démontre ne pas être aléatoires, nous permet d'inférer les lois qui guideraient le tirage. Jusque là, cette vision est proche de celle qui avait été adoptée par Shannon dans sa formulation de la théorie de l'information.
Comme en cryptographie ou dans certaines approches neuroniques, nous nous intéressons aux répartitions fréquentielles des lexèmes (chaînes de caractères obtenues). Ceci fournit d'abord des résultats intéressants dans l'esprit de l'approche lexicométrique traditionnelle.

Mais à partir de là, notre heuristique diverge lorsqu'il s'agit d'analyser le *contenu* des textes. En effet, nous pouvons renoncer à l'approche statistique naïve puisque le code sémantique utilisé nous est connu : nous parlons la langue naturelle, nous partageons à peu près la même vision du monde que les auteurs des textes analysés, et ces textes font référence à ce monde. La signification des mots nous est connue, contrairement au cryptographe. Ce qui nous intéresse, c'est la façon dont les textes assemblent ces mots pour créer des noyaux de sens, combinaisons particulières qui font référence à un tableau du monde donné.

Un noyau de sens donné peut se présenter dans les textes sous des formes différentes, en raison de la souplesse de la langue, et aussi en raison de la nature relativement floue des tableaux représentés par les combinaisons de mots.

Par exemple :

```
Belle marquise, vos yeux me font mourir d'amour
Belle marquise, d'amour vos yeux mourir me font
Je meurs d'amour pour les beaux yeux de la marquise
J'aime la marquise, ses beaux yeux m'ont séduit
```

sont des formes différentes d'un même noyau de sens.

Nous pouvons alors rechercher des *paradigmes*, qui sont l'ensemble des combinaisons de traits (ici, des mots) qui commutent dans l'expression d'une même représentation. Cette approche a été exposée en détail (Lahlou, 1992).

Avec le recul, il nous est apparu que l'application de la méthode au langage peut être considérée comme un cas particulier d'un principe plus général qui peut être appliqué à la reconnaissance de formes à partir d'une description quelconque en langage libre. C'est donc une théorie générale de l'interprétation sémantique des corpus que nous proposons ici.

1. ANALYSE LEXICALE, ANALOGIE ET COMPRÉHENSION DU MONDE

> "(...) ce monde fut créé par combinaison de possibles qui, dans l'entendement divin, forment éternellement les notes en ségrégation (*sejuncta*), lettres ou nombres, d'une table qui convient à la nôtre en certains rapports. (...)
> Or, nous trouvons tous ces éléments «pêle-mêle», multitude de «détails» infiniment complexe, qui répond originairement à ce que le phénomène mondial a de «multiplex». Science des complexions élémentaires, la combinatoire est, pour Dieu, science des possibles et organon de la constitution du monde, elle est, pour nous, doctrine de déchiffrement de l'univers (...)".
>
> (Serres, 1968, pp. 105-107, passim)

1.1. Descriptions du monde, construction des formes et des catégories

1.1.1. Descriptions du monde et construction des formes

Le monde *décrit* (par un système formel quelconque) se présente sous la forme d'une combinatoire de signes ou alphabet de l'observateur. L'expression de cette combinatoire obéit à des règles d'énonciation, conventions formelles qui régissent l'assemblage de surface des signes. Par exemple, dans la langue naturelle, la syntaxe, la grammaire, la ponctuation ; dans la représentation graphique, les conventions d'échelle métrique. Ces conventions permettent au destinataire de la description de reconstruire une certaine image de la structure de ce qui est décrit ; il s'agit d'un implicite mode d'emploi procédural pour reconstruire le référent (l'objet représenté) à partir du discours (le texte). Le référent est interprété/reconstruit à partir du signe en passant par deux niveaux de construction :

- l'un sémantique : la langue, système encyclopédique de références au monde, qui *constitue* le sens ;
- l'autre formel : le code technique qui règle la formation de *l'expression* du sens par l'assemblage de signes (mots, caractères).

L'ensemble constitue un système symbolique ; le résultat tangible de l'utilisation de ce système par le scripteur (ou le locuteur) produit des textes qui sont sensés exprimer dans leur forme et leur structure des formulations du monde (descriptions, propriétés, assertions, etc.). Ce sont bien ces dernières qui nous intéressent ; de notre point de vue toutes les règles formelles sont alors des artefacts, des parasites qu'il nous faut filtrer lors de l'analyse.

Cette vision est valable pour d'autres systèmes sémantiques que la langue naturelle, par exemple les fichiers statistiques. Dans un fichier d'enquête, la position d'un code donné permet d'interpréter la signification du code : "04" en colonne 12 de la ligne 30 signifie que pour l'individu n°30 la modalité de la variable 12 est "04", soit encore que Paul Dupons

(individu 30), à la question "Comment allez vous au bureau ?" (variable 12) a répondu "en autobus" (modalité 04). La compréhension du sens de ce code (Monsieur Dupons va au bureau en autobus), fait appel à une connaissance implicite du monde de référence :

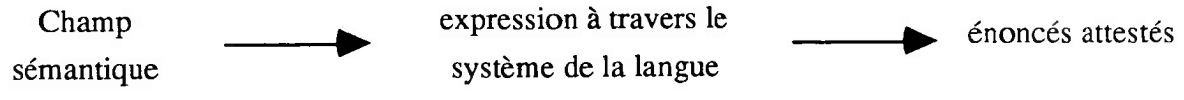
- Dupons est un homme ;
- un homme travaille (il a des activités régulières dans un lieu nommé bureau) ;
- un homme habite dans un lieu distinct de son bureau ;
- il doit donc se déplacer de ce lieu à son bureau pour travailler ;
- un autobus est un moyen de déplacement ;
- etc.

Le fait que la modalité soit 4 est une convention de codage, elle aurait bien pu être 5 avec un autre système de codage. Lors de l'analyse, il faudra veiller à ce que ces conventions restent neutres, par exemple il faudra éviter des traitements dans lesquels la modalité 4 serait perçue comme "plus grande que" la modalité 3, car cette différence n'a pas de sens dans notre perspective. Si ce codage avait une influence, ce serait un artefact.

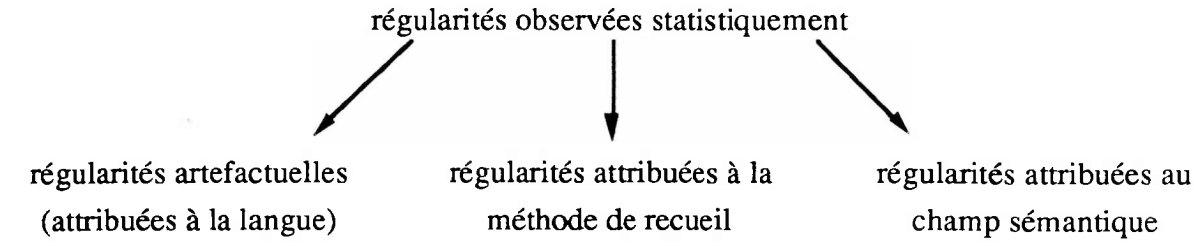
Les règles d'énonciation formelles sont *internes* au système de description. Elles sont arbitraires et contingentes, et provoquent dans les descriptions observées des régularités *artefactuelles* dans ce qu'on appelle le niveau de surface de l'expression (sa forme physique, en faisant abstraction du réseau sémantique sous-jacent).

Il existe dans les descriptions du monde observable d'autres régularités, qui ne sont pas réductibles aux règles d'énonciation formelles : ce sont ces régularités que nous attribuons à la structure naturelle du monde.

L'hypothèse implicite que nous faisons du mécanisme de production d'énoncés linguistiques est la suivante. Il existe un champ sémantique, sorte de réseau de concepts. Chacun de ces concepts est exprimable sous la forme d'un paradigme, classe d'expressions formellement différentes qui renvoient toutes (à peu près) au même sens.



L'analyste est confronté au résultat de ces énoncés. Il constate plusieurs types de régularités, dont il attribue certaines à la forme du champ sémantique.



Cette façon de raisonner est saine sur le fond ; en fait elle est inapplicable rigoureusement, car la description n'est jamais vraiment indépendante du système de description. C'est donc un idéal qui ne peut être atteint. Avant d'aller plus loin, précisons pourquoi et nous ne reviendrons plus sur cette question.

1.1.2. L'effet pochoir

Chaque description du monde, en tant qu'elle est une représentation de quelque chose, cherche à traduire et transmettre une forme, celle qu'a observée l'auteur "D" de la description. Cette forme est, par construction, une interprétation par D de son observable, elle n'est qu'une vision partielle et idiosyncrasique de ce qui est décrit, elle comporte une part indéterminable de subjectivité. Cette part est *absolument* indéterminable au sens de Heisenberg, car toute observation est expérience, c'est-à-dire processus singulier dont l'observateur pourtant définit arbitrairement une part qui sera "le phénomène", et l'autre "le dispositif d'observation". En termes gestaltistes : il est impossible de décrire une forme sans le fond, mais celui-ci reste indescriptible. Seul un consensus entre observateurs peut définir le fond. Dans la pratique, on procède en demandant à chaque observateur d'utiliser les mêmes règles d'énonciation, et les régularités qui sont indépendantes de l'expérience définissent alors ce fond, par implicite.

Il ne s'agit pas spécifiquement d'un problème d'analyse lexicale, mais d'un problème épistémologique général. Ainsi, dans l'étude de la cognition, la difficulté de démêler la "source" de son "effet" est apparue tellement insoluble que certains éthologistes en sont venus à adopter la notion ambivalente de "stimulus-contrast" dans laquelle on ne tranche pas entre le fait que le stimulus est défini par contraste avec son contexte, et le fait qu'il déclenche une réponse qui est, elle-même, définie par contraste avec son contexte comportemental.

Le terme de "stimulus-contrast" vient de Andrew (1963). La critique de l'ambivalence de ce terme se trouve dans une note de Hinde R.A. (1971). Le problème de l'ambivalence du terme a empêché la traduction du terme dans l'édition française d'un article de Van Hoof J.A.R.A.M. (1978) comme en témoigne une note du traducteur p. 21. L'histoire de ce terme n'est pas

anecdotique : l'éthologie est probablement, avec la physique, et la psychanalyse, la discipline dans laquelle la réflexion sur l'acte d'observation a été le plus poussée. On tombe ici sur une question fondamentale : le message ne peut pas être indépendant du code, et cette difficulté apparaît même dans des circonstances où l'on cherche à décrire des phénomènes "simples".

> "Tout élément dans la série est simultanément stimulus, réponse, et renforcement. Soit un élément de comportement donné imputable à A. Il constitue un stimulus dans la mesure où il est suivi d'un élément imputable à B, et où celui-ci est à son tour suivi d'un élément imputable à A. Mais, pour autant qu'un élément de A est pris en sandwich entre deux éléments de B, il constitue une réponse. De même, il est un renforcement dans la mesure où il suit un élément de B. Les échanges que nous considérons ici constituent donc une chaîne formée de maillons triadiques qui se chevauchent, et dont chacun est comparable à une série stimulus-réponse-renforcement."
>
> (Bateson, Jackson ,1964)

Trancher la question sur le plan théorique est, par construction impossible : car le signe exprimé renvoie à la fois à un référent et à un système arbitraire et conventionnel de perception et de description qui sépare de façon contingente le phénomène du contexte :

> "La formation des concepts commence avec la perception de la forme. Alors que l'image optique projetée sur la rétine est un enregistrement mécaniquement complet de son homologue matériel, le percept visuel correspondant ne l'est pas. Percevoir la forme consiste à appréhender les caractéristiques structurales que contient le matériau-stimulus ou qui lui sont conférées. Il est rare que ce matériau soit exactement conforme aux formes qu'il acquiert dans la perception. (...) L'important, c'est qu'on puisse dire d'un objet regardé par quelqu'un qu'il n'est vraiment perçu que dans la mesure où il s'adapte à une forme organisée. (...) Strictement parlant, un percept ne se réfère jamais à une forme unique, individuelle ; il renvoie plutôt à un genre de modèle dont relève le percept. Il peut n'y avoir qu'un seul objet correspondant à ce modèle, comme il peut y en avoir d'innombrables. (...) Il n'y a donc aucune différence entre percept et concept, ce qui est tout à fait conforme à la fonction biologique de la perception. Pour être utile, la perception doit renseigner sur les types des choses sans quoi les organismes seraient incapables de mettre à profit l'expérience vécue."
>
> (Arnheim, 1976)

Dans le cas de la langue, le contexte explicatif des faits décrits se situe pour partie dans le co-texte, et pour partie dans le réseau de la langue elle-même.

> "Le signe (la phrase) prend un sens par référence à un système de signes ou langage, auquel il appartient. Comprendre une phrase, c'est en somme comprendre un langage."
>
> (Wittgenstein, 1965)

Whorf (1956) a bien montré dans le cas des langues naturelles que ceci n'est pas une simple argutie théorique. La langue naturelle découpe artificiellement le tissu continu du monde en concepts, et en mots. Ce découpage, comme la langue est un système pragmatique, a une influence réelle sur les comportements et les perceptions. Or, ce découpage ne correspond pas forcément à des universaux de la pensée. Ainsi, la langue Hopi distingue des temps, des modes, et des structures d'action qui sont très différentes de ceux du français, par exemple. Il

y a là ce que nous pourrions appeler un effet de pochoir ; l'observateur ne peut percevoir (et décrire) le phénomène qu'à travers le filtre de son système d'observation et d'expression.

> "Tous les observateurs ne sont pas conduits à tirer, d'une même évidence physique, la même image de l'univers, à moins que l'arrière-plan linguistique de leur pensée ne soit similaire, ou ne puisse être rendu similaire d'une manière ou de l'autre."[15]
> "le langage est [donc] avant tout une classification et une réorganisation opérées sur le flux ininterrompu de l'expérience sensible, classification et réorganisation qui ont pour résultat une ordonnance particulière du monde..."[16]
> "Chaque langage est un vaste système de structures, différent de celui des autres langues, dans lequel sont ordonnées culturellement les formules et les catégories par lesquelles l'individu non seulement communique, mais aussi analyse la nature, aperçoit ou néglige tel ou tel type de phénomènes ou de relations, et dans lesquelles il coule sa façon de raisonner, et par lesquelles il construit l'édifice de sa connaissance du monde"[17].

Cette remarque inter-culturelle est également valable à l'intérieur d'une même langue entre différents utilisateurs. La taille du vocabulaire disponible, et des concepts sous-jacents, limite la conceptualisation et l'expression à la *culture* du locuteur.

Ce problème est insoluble. Nous le mentionnons pour qu'on reste conscient que, de toutes façons, l'expression linguistique d'un objet est déjà entachée de biais au départ. Il faut donc raison garder lorsque les méthodes de traitement ajoutent d'autres biais ; car elles ne biaisent pas un objet parfait et sacré, "l'observation originale", elles ne font qu'ajouter des biais à une expression déjà imparfaite.

Il ne faut donc pas hésiter outre mesure à agir sur le texte, sous le prétexte qu'on abîmerait le corpus de manière sacrilège.

1.1.3. Construction des catégories

Comme il n'existe pas de solution absolue pour déterminer les contours des formes des phénomènes, il ne reste que la possibilité de chercher une solution idoine (adaptée aux objectifs pragmatiques).

On procède dans l'approche scientifique à une reconnaissance empirique fondée sur la recherche de ce qui est analogue, ou plus exactement système-compatible (Lahlou, 1990) entre les différentes descriptions, pour déterminer les grandes constantes formelles des phénomènes (en particulier, des objets linguistiques). Faute de pouvoir trouver *la* solution, on

---

[15]Benjamen Lee Whorf, Language, p. 214, cité par Mounin, p. 46.
[16]Whorf, Language, p. 55 cité par Mounin, p 46.
[17]Whorf, Language, p. 239, cité par Mounin p. 47.

s'accorde sur un consensus, en confrontant les différentes subjectivités pour trouver un compromis acceptable. Nous présentons ici un modèle de ce processus d'invention scientifique, dans des termes qui nous permettent de mieux expliciter notre démarche dans le domaine particulier de la recherche de représentations à partir de l'analyse des textes.

La détermination des formes se fait donc, dans les approches scientifiques comme dans les approches naïves, par une comparaison d'énoncés descriptifs dont on extrait les régularités. Celles-ci sont, comme nous l'avons vu, de trois types :

- les régularités artefactuelles, facilement détectables car elles sont construites par des conventions formelles explicites (celle du système de description) ;
- les régularités indépendantes de l'expérience, et donc attribuées au système d'observation (observateur inclus). Celles-ci comprennent d'ailleurs les précédentes ;
- les autres régularités. Ce sont ces dernières qui sont attribuées au phénomène observé, et que nous appellerons la forme du phénomène.

A ces régularités, on cherche alors des règles de production économiques en termes formels : ce sont celles-ci qui sont les *explications*. Par exemple, en observant les trajectoires de solides en chute libre, on peut trouver des régularités entre les mesures de temps et de distances parcourues. L'équation qui permet de produire (et donc, de reproduire) ces régularités est appelée loi et est considérée comme l'explication du phénomène observé. Les critères de validation (prédiction, comparabilité, falsifiabilité, ...) sont simplement des vérifications du caractère "réellement" productif des règles.

Or, nous l'avons vu, les énoncés descriptifs du monde se présentent forcément sous la forme de combinatoire des signes conventionnels. On peut alors, pour résumer le processus de recherche des formes et de leurs explications, le décrire de la façon suivante :

1) on obtient, par une série d'expériences, des vues subjectives et partielles exprimées sous la forme de combinaisons de signes conventionnels ;

2) on compare ces expressions pour en tirer des régularités caractéristiques du phénomène, c'est-à-dire des combinaisons de signes *analogues* (nous reviendrons sur ce point) ;

3) on construit des règles de production de ces combinaisons (règles d'assemblages des signes conventionnels) ;

4) on vérifie que les productions de ces règles correspondent effectivement à des observables, par des expériences de validation.

Que sont des combinaisons *analogues* ? Ce sont des combinaisons identiques d'un certain point de vue. C'est dire soit :

a) que, par une transformation licite selon les conventions d'expression, elles sont identiques. C'est le sens le plus strict :

> ANALOGIE ◊ 2. (1520, en math.). SC. Identité de deux rapports. > Proportion, proportionnalité. Le principe d'analogie. Termes (grand, moyen, petit terme) d'une analogie. Analogie arithmétique, géométrique. L'équation a = c est une analogie. - Raisonnement par analogie, qui, étant donné trois termes d'une analogie, détermine le quatrième.
>
> Le Robert Électronique

Par exemple, en remplaçant les occurrences de signes de l'expression par des catégories (taxinomiques, fonctionnelles...) qui les incluent, on obtient des expressions strictement identiques. Ainsi, les descriptions suivantes sont analogues :

la vache broute le gazon -(hyp)-> *le bovin se nourrit de l'herbacée*
le boeuf mange l'herbe -(hyp)-> *le bovin se nourrit de l'herbacée*

comme on s'en rend compte en remplaçant les mots par leurs hyperonymes.

Il existe différents niveaux d'analogie, qui fournissent des règles plus ou moins générales selon la taille des catégories référées. Ainsi les deux descriptions précédentes sont analogues à :

Paul Dupons digère son cassoulet

par la transformation :

l'être transforme la substance.

b) que les deux descriptions contiennent une sous-partie identique. C'est le sens courant, le plus large :

> ANALOGIE ◊ 1. Cour. Ressemblance établie par l'imagination (souvent consacrée dans le langage par les diverses acceptions d'un même mot) entre deux ou plusieurs objets de pensée essentiellement différents. > Affinité, association, correspondance, lien, parenté, rapport, relation, similitude. (…) Conclure de l'analogie des causes à celle des effets, de l'analogie des parties à celle du tout, de l'analogie des moyens à celle de la fin. Juger, suggérer, évoquer qqch. par analogie. (…)
> Philos. et cour. Jugement, raisonnement par analogie, qui conclut d'une ressemblance partielle à une autre ressemblance plus générale ou totale. > Induction. (…)
> Hist. ling. Régularité systématique (de la langue), opposé à l'anomalie* (de l'usage).
> Rapport existant entre mots appartenant à un même champ sémantique.
>
> Le Robert Électronique

La recherche de combinaisons de signes analogues est un moyen de rechercher au-delà des régularités artefactuelles, les régularités inhérentes au phénomène. Dans cette perspective, l'analogie permet de repérer des formes par induction, ou par abduction.

> INDUCTION ◊ 2. (XIVe). Log. Opération mentale qui consiste à remonter des faits à la loi, de cas donnés (propositions inductrices) le plus souvent singuliers ou spéciaux, à une proposition plus générale. > Analyse, généralisation (- Image, cit. 48). Induction formelle ou par énumération. Induction amplifiante, qui étend à tout un genre ce qui a été constaté dans un certain nombre de cas singuliers (syn. : induction baconienne).
> Le fait de remonter par le raisonnement ou l'intuition, de certains indices à des faits qu'ils rendent plus ou moins probables. L'induction est un processus de pensée reconstructif. > Inférence. Raisonnement par induction. > Analogie (cit. 7), association (des idées). Dégager par induction des vérités implicites*.
>
> ABDUCTION ◊ 2. (1771). Log. Syllogisme dans lequel la majeure est certaine et la mineure probable (la conclusion est alors seulement probable).
> Le Robert Électronique

On suppose que les différentes descriptions empiriques sont des formulations incomplètes, et éventuellement partiellement inexactes, d'un type idéal constitué d'un assemblage de traits, déduit par induction des combinaisons descriptives observées.

Sur le plan linguistique, ce type d'opération se traduit par des figures de rhétorique qui visent à accentuer l'efficacité descriptive tout en conservant une forme synthétique, dont la plus connue est la métaphore[18], et qu'on appelle les *tropes*.

> Les tropes sont des figures par lesquelles on fait prendre à un mot une signification, qui n'est pas précisément la signification propre de ce mot (...) Ces figures sont appelées tropes du grec tropos (...) dont la racine est trepo, verto, je tourne.
> (C. Du Marsais, Des tropes..., I, IV.)

D'autres figures (métonymie, synecdoque) utilisent ce mécanisme de "la partie pour le tout" qui semble être une caractéristique du fonctionnement cognitif humain, consistant à classer ensemble les objets "qui se ressemblent" à partir de la reconnaissance de traits identiques. Nous avons d'ailleurs utilisé ce mécanisme pour construire, concrètement, des systèmes de codification assistée par ordinateur (Lahlou, Maffre et Beaudouin, 1991 ; Lahlou, 1992 b). On se reportera avec profit au remarquable ouvrage de Michel Foucault (1966), Les mots et les choses, pour une discussion détaillée de cette question à travers l'histoire de la philosophie occidentale, et aux travaux classiques d'Eleanor Rosch en Psychologie Cognitive sur l'approche prototype /qualité (Rosch, 1975, 1976).

---

[18] MÉTAPHORE ◊ Figure* de rhétorique, procédé de langage qui consiste dans un transfert de sens (terme concret dans un contexte abstrait) par substitution analogique. (Le Robert).

1.2. Construction empirique de catégories par analyse lexicale

Ce cadre heuristique général étant posé, voyons-en les applications à notre problème particulier.
Notre objectif est de repérer des représentations à partir de corpus linguistiques. Dans la pratique, nous allons réaliser les différentes opérations de repérage des formes de la façon suivante :

1) Rassembler un corpus de textes "à propos de" l'objet qui nous intéresse, par enquête, ou sélection dans des bases de données. *(On obtient, par une série d'expériences, des vues subjectives et partielles exprimées sous la forme de combinaisons de signes conventionnels)*

2) Réaliser une classification de ces textes dans l'espace de leur expression lexicale, en rapprochant, par des méthodes statistiques, ceux qui ont des traits linguistiques identiques -mots ou chaînes de mots en commun- *(on compare ces expressions pour en tirer des régularités, c'est-à-dire des combinaisons de signes analogues) ;*

3) Interpréter les classes en termes sémantiques. Soit :

a) Repérer les combinaisons de traits typiques de chaque classe ;

b) Construire, dans l'espace sémantique, un noyau de sens, constitué d'une combinaison de traits sémantiques induite de l'observation des traits typiques de la classe ;

c) Formuler des lois d'expression de cette combinaison sémantique dans la langue.

Pratiquement, on considérera que les règles de production sont l'ensemble des règles de la langue appliquées au noyau sémantique en question. Autrement dit, toutes les expressions "bien formées" du noyau sémantique sont des productions acceptables.*(On construit des règles de production de ces combinaisons (règles d'assemblages des signes conventionnels)) ;*

4) Vérifier que les expressions effectivement classées dans chaque classe sont interprétables comme expression de ces formes mentales sous-jacentes, en fonction des règles d'expression fixées par la langue, et des relations sémantiques sur le monde issues non seulement des associations attestées du corpus, mais de la connaissance du sens commun sur le monde *(on vérifie que les productions de ces règles correspondent effectivement à des observables, par des expériences de validation).*

La principale différence avec la recherche en sciences exactes réside dans le point 3. En effet, comme on le constate, les traits qui sont utilisés pour construire nos lois de production ne sont pas directement calculables à partir des traits manifestes, ils sont inférés. En physique, les paramètres observables rentrent directement dans la formulation des lois. Pour nous, les objets pertinents sont des concepts ou des noyaux de sens, qui ne sont pas de la même catégorie ontologique que les observables (mots). Et nous ne pouvons pas désigner simplement les variables et les constantes dans le formalisme du sens commun. En mathématiques, on peut désigner des variables (par exemple la masse) et des mesures de ces variables. Il existe une relation simple entre les termes : une variable s'exprime par une mesure dans la réalité. Pour notre problème, la variable n'est pas un outil théorique pertinent, c'est de "concepts" ou de "noyaux de sens" que nous sommes occupés. Or ces concepts s'expriment non pas par un seul trait (ou variable) mais par une combinaison de traits. Ce qui est "mesure" dans les sciences physiques est ici remplacé par une expérience (équivalent du processus de mesure) *et* une interprétation sémantique. Dans l'analyse lexicale, le processus de mesure véhicule donc une plus grande quantité de connaissances implicites.
Notre méthode vise à instaurer un processus de catégorisation qui prenne en charge automatiquement une grande partie des opérations de ce "calcul analogique" qui fait référence à la connaissance implicite du monde, sans utiliser cette dernière car elle est techniquement lourde à manipuler.

La méthode de classification utilisée fonctionne par comparaison des traits composant les textes, elle est donc essentiellement tropique (on infère de l'analogie des parties l'analogie du tout). Le principe qui guide l'interprétation est abductif :

Lorsqu'on compare plusieurs expressions, on constate que :

(1) - plusieurs expressions (Xi) possèdent le trait "a"

On en induit :

(2) - il existe *probablement* une classe A dont "a" est un trait caractéristique

soit : il est probable que (si $Xi \supset a$, alors $Xi \in A$)

On abduit de (1) et (2) :

(3) - alors Xi fait partie de la classe A.

La classification descendante construit d'un même mouvement les traits caractéristiques et les classes, car la probabilité que "a" puisse être un trait provient précisément de ce que "a" est observé sur un certain nombre d'expressions. En d'autres termes, on choisit comme traits classificatoires les traits qui sont effectivement susceptibles d'être productifs de classifications

intéressantes, c'est-à-dire qui permettraient de caractériser un nombre suffisant d'expressions à classer. Les traits trop répandus (non discriminants) ou les traits trop rares (hapax, traits aberrants) ne sont pas intéressants. C'est là une résolution empirique du problème des critères pertinents.

Par exemple, dans la description de 1 000 petits déjeuners, le mot *noir* est assez fréquent, il intervient fréquemment associé à *café* ("café noir"). Sa situation assez fréquente, mais pas trop, fait qu'il a contribué à la formation d'une classe (Beaudouin et al., juin 1993) :

> **Café noir (78 individus)**
> Cette classe est caractérisée par l'adjectif *noir*, et même par l'expression *café noir*. *Noir* est bien sûr la plupart du temps associé à *café*, mais on constate quelques cas où il est accolé à un autre mot : *crêpes au blé noir*, *chocolat noir en morceaux*, *pain noir*. Généralement, les aliments solides pris avec le café noir sont peu nombreux et ce sont surtout du pain, grillé ou non, tartiné de beurre et de confiture. 14 réponses (22 %) ne comportent que le café noir ; 35 réponses (55 %) ne citent comme aliment solide que du pain, du beurre et de la confiture. Pour finir, citons quelques réponses caractéristiques de cette classe :
> "café noir ";
> "café noir, croissant ou biscottes" ;
> "café noir, Canderel" ;
>
> (Beaudouin et al., juin 1993)

Cette méthode de classement est imparfaite dans le détail ; le trait *noir* a fait classer dans cette même catégorie des réponses comme *pain noir*, qui ne correspondraient peut-être pas à nos critères sémantiques de classement. Par ailleurs, comme ce trait (noir) rentre en compétition avec d'autres traits pour classer les réponses, certaines qui comportaient effectivement le mot *noir* et d'autres traits classificatoires ont été classées dans d'autres classes que la classe "café noir". Cet exemple est caricatural, mais il illustre une difficulté générale de la classification. Car, en fait (et y compris dans notre logiciel) tous les traits sont a priori pertinents. Mais certains le sont plus que d'autres, et l'arbitrage, en l'absence d'une connaissance préalable des "vraies" catégories, ne peut être fait que sur des critères statistiques. C'est-à-dire que l'on favorise les traits qui sont susceptibles de donner une bonne partition d'un point de vue statistique (ce qui ne veut pas dire la partition juste).

Cette procédure empirique, qui sélectionne les traits au vu de leur pertinence classificatoire dans le corpus même, résout la difficulté des classifications a priori, où les traits, par leur choix, leur recouvrement, ou leur niveau de détail, ne permettent pas d'obtenir une partition harmonieuse des objets à classer. Elle évite des classifications du type de celle que Borges (1957, p. 144) fait attribuer par le docteur Franz Kuhn à "une certaine encyclopédie chinoise intitulée "Le marché céleste des connaissances bénévoles" où :

> "les animaux se divisent en a) appartenant à l'Empereur, b) embaumés, c) apprivoisés, d) cochons de lait, e) sirènes, f) fabuleux, g) chiens en liberté, h) inclus dans la présente classification, i) qui s'agitent comme

des fous, j) innombrables, k) dessinés avec un très fin pinceau de poils de chameau, l) et coetra, m) qui viennent de casser la cruche, n) qui de loin semblent des mouches".

et dont l'effet comique provient de la non pertinence *vis-à-vis du corpus des animaux considérés dans son ensemble* des traits classificatoires choisis, qui sont des traits locaux.

La procédure a cependant l'inconvénient de donner des traits fortement dépendants du corpus utilisé, ce qui peut être embarrassant pour faire des comparaisons entre corpus.

Cette approche est donc purement exploratoire et a pour objet les unités de contexte qui sont, en quelque sorte, des "présomptions de paradigme". C'est à ces objets que nous appliquons la méthode de classement analogique abductif décrite plus haut.

Pratiquement, au niveau sémantique, on découvre que différentes UCE sont analogues (car elles ont un contenu lexical en commun). On les met dans une même classe (inférence). On considère que cette classe est caractérisée par le contenu lexical commun (abduction).
Ce contenu lexical est alors interprété, à partir de notre connaissance du monde, comme un paradigme. En fait, la combinaison des mots les plus typiques de la classe sert à construire le paradigme (combinaison typique = prototype, on considère que le prototype est l'expression canonique d'un paradigme). Ce paradigme est en lui-même une loi de production, en ce qu'il est la classe de toutes les expressions syntaxiques de ses commutants sémantiques, par homologie (niveau sémantique), allotaxie (niveau syntaxique) et allographie (niveau lexical).

Pour prendre un exemple concret, la classe "petit déjeuner classique" interprétée à partir du contenu lexical commun des UCE qu'elle contient, est considérée comme le paradigme : absorption conjointe de boisson chaude et de solides panifiés tartinés. Elle peut s'exprimer au niveau homologique par la combinaison des homologues sémantiques de chacun de ses termes.

manger	café	pain
boire	thé	tartine
absorber	boisson	...
prendre	...	...
...	...	...

Au niveau syntaxique et lexical, toutes les variations autorisées par la langue naturelle sont des expressions admissibles (productions conformes au paradigme).

Exemple :

```
Je prends du café et des tartines
Café au lait, pain beurré
```

L'analyse fine démontre qu'il existe d'ailleurs une grammaire locale qui régit les énonciations admissibles. Par exemple, dans le cas du petit déjeuner "classique", nous avons pu mettre en évidence la grammaire locale suivante :

> Il nous a semblé, au vu des analyses effectuées grâce à ALCESTE sur les produits consommés tant la semaine que le week-end, que les réponses avaient une structure (ou grammaire) bien particulière que l'on pourrait appeler grammaire du petit déjeuner :
> **Petit Déjeuner = boisson chaude + {solide panifié} + {autres produits}**
> Les éléments entre accolades sont facultatifs.
> **boisson chaude** = boisson chaude + {précisions boisson}
> boisson chaude = café, thé, chocolat, lait, Ricorée, chicorée...
> précision boisson = au lait, sucré, sans sucre, chaude, avec du citron...
> **solide panifié** = solide panifié + {précisions solide}
> solide panifié = pain, biscottes, croissant, tartine, gâteaux, biscuit, baguette...
> précision solide = grillé, avec du beurre, avec de la confiture, frais, de la veille...
> **autres produits** = jus d'orange, céréales, yaourts, fromage...
> Par exemple, *café noir, une biscotte beurrée* suit la grammaire.
> **boisson chaude** = café {noir}
> **solide panifié** = biscotte {beurrée}
> **autres produits** = Ø
> De même que café au lait, pain frais, beurre, confiture, un yaourt.
> **boisson chaude** = café {au lait}
> **solide panifié** = pain {frais, beurre, confiture}
> **autres produits** = yaourt.
> Par contre, café sans sucre, un demi pamplemousse ne suit pas la grammaire.
>
> (Beaudouin et al., juin 1993)

Pour finir, soulignons qu'il ne faut pas accorder une importance excessive aux graphes de mots finalement obtenus, puisque les positions relatives sont déterminées essentiellement par des artefacts de méthode. Cette méthode, à laquelle il faut conserver une valeur *uniquement exploratoire* présente néanmoins l'avantage d'être opératoire, et de faire sens. Les interprétations et les théories éventuelles ne sauraient en aucun cas résulter de la méthode elle-même, mais des connaissances pragmatiques de l'analyste. Cette mise en garde semble une évidence, cependant nous ne saurions la rappeler avec trop de vigueur, car l'expérience montre que l'aspect séduisant des représentations graphiques, et le caractère fortement projectif que leur confère le fait que chaque point représente un mot, est une tentation permanente pour l'analyste à considérer que ses interprétations ne sont qu'une explicitation d'une structure révélée par l'analyse.

Il résulte de cet état de fait que ce type de méthode ne devrait pas être utilisé sans supervision d'un praticien rompu à l'analyse statistique, afin de garder constamment à l'esprit les artefacts

de méthode sous-jacents. Il est à craindre que la facilité d'utilisation des logiciels, et la facilité apparente de l'interprétation ne suscitent une foule d'analyses abusives par des amateurs. C'est précisément ce problème qui a jeté le discrédit sur les méthodes d'analyse factorielle à la française, à la suite de leur utilisation à tort et à travers par des analystes insuffisamment formés et avides de "résultats". Rappelons qu'un listing n'a jamais été une analyse à lui seul, et encore moins un "mapping".

1.3. Représentation et monde

Ce sont principalement des représentations sociales que nous cherchons à repérer à travers les analyses de corpus. Comment se fait le passage des "classes" observées aux représentations ?

Il est ici nécessaire de faire le point sur la théorie des représentations mentales et sociales sous-jacente à notre méthode d'analyse. Ces questions ont été décrites en détail sur un plan technique dans des travaux précédents (Lahlou, 1990, Lahlou, 1992) ; on fournit ici un aperçu qui en situe les grands principes à partir du système du monde de Spinoza, dont l'expression est non seulement antérieure mais aussi plus heureuse et plus claire que la nôtre.

L'idée est simple. Les représentations sociales sont des outils de pensée à visée pragmatique, c'est-à-dire qu'ils permettent à l'individu de penser le monde pour agir dessus. Ce sont des représentations *de* quelque chose *pour* celui qui pense (Moscovici, 1976 ; Jodelet, 1984, 1989). Ces représentations, en même temps que le langage qui les exprime, sont apprises par l'expérience concrète du monde au cours de la vie de l'individu. Chaque situation vécue *associe* des perceptions des objets du monde ; la mémoire conserve ces associations. Se constitue ainsi un vaste réseau d'associations mentales qui relie entre eux les images des objets[19], et le vécu des situations correspondantes (motivations, sensations, émotions, motricité...). L'aspect pragmatique provient de ce que la structure d'enregistrement possède une interface avec le monde. Elle est aussi susceptible de fonctionner (pour utiliser une image mécaniste) comme structure de lecture, et donc de produire des réponses motrices. C'est ce qu'on désigne par "apprentissage".

[19](...) nous appellerons «images des choses» les modifications du corps humain dont les idées nous représentent les corps extérieurs comme présents pour nous, malgré que ces images ne reproduisent pas la configuration exacte des choses, et lorsque l'âme se représente les choses par ce procédé, nous dirons qu'elle «s'imagine» (Spinoza, Éthique, livre II, scholie du théorème XVII).

L'ensemble des représentations est donc une image subjective du monde. Cette image est homéomorphe au monde, du moins d'un certain point de vue, elle en conserve la structure. *Par conséquent, lorsque nous étudierons l'enchaînement des idées (par exemple exprimé à travers nos corpus), nous retrouverons quelque chose de l'enchaînement du monde tel qu'il est vécu par le sujet.*

Le raisonnement de Spinoza l'explicite bien :

> L'ordre et l'enchaînement des idées est le même que l'ordre et l'enchaînement des choses (Éthique, livre II, théorème VII)

car :

> Comme les pensées et les idées des choses s'ordonnent et s'enchaînent dans l'Âme, de même exactement les affections du Corps et les images des choses s'ordonnent et s'enchaînent dans le corps. (Éthique, livre V, théorème I).

car :

> Si le corps humain s'est trouvé une fois affecté simultanément par deux ou plusieurs corps, dès que, dans la suite, l'âme s'imaginera l'un quelconque d'entre eux, elle se souviendra aussitôt des autres. (Éthique, livre II, théorème XVIII)
> Scholie : Par là nous comprenons immédiatement ce qu'est *la mémoire*. Ce n'est pas en effet autre chose qu'un certain enchaînement d'idées qui enveloppent en elles la nature de choses existantes en dehors du corps humain, enchaînement qui se produit dans l'âme, suivant l'ordre et l'enchaînement des modifications du corps humain. (...)
> Ainsi par exemple, de la pensée du mot *pomus* un Romain tombera immédiatement dans l'idée d'un fruit qui n'a aucune ressemblance avec ce son articulé, ni avec lui rien de commun, si ce n'est que le corps de cet homme a souvent été affecté par ces deux choses à la fois ; c'est-à-dire que ce même homme a souvent entendu le mot *pomus* en même temps qu'il voyait le fruit. Ainsi chacun tombera d'une pensée dans une autre suivant que, dans le corps de chacun, l'habitude aura ordonné les images des choses. Car un soldat, par exemple, voit dans le sable les traces d'un cheval, aussitôt, de la pensée du cheval il tombera dans celle du cavalier, de là dans la pensée de la guerre, etc. Le paysan, au contraire, tombera de la pensée du cheval, dans celle de la charrue, du champ, etc. Ainsi chacun, suivant qu'il aura contracté l'habitude d'associer et d'enchaîner les images de telle ou telle façon, tombera d'une pensée donnée dans une autre.
>
> (Éthique, livre II, scholie du théorème XVIII).

Cette vision, qui justifie directement notre approche par *associations lexicales*, est étonnamment moderne, très proche de l'approche cognitiviste par réseaux neuronaux, qui cherche à reproduire les mécanismes d'apprentissage décrits notamment par Pavlov et Skinner, en construisant des "boîtes noires" qui connectent stimulus et réponse. Dans une telle perspective, les "idées" qui nous intéressent -et que nous cherchons à reconnaître par analyse des corpus- peuvent être considérées comme des ensembles fonctionnels créés par associations d'expériences concrètes du corps.

Ces noyaux associatifs réagissent les uns sur les autres à des niveaux plus ou moins agrégés, à l'intérieur de la boîte noire qu'est le cerveau (version moderne de l'âme de Spinoza). Minsky les appelle des "simulus" (stimuli internes) :

"Quel type de processus pourrait reproduire l'image d'une présence imaginaire ? Les scientifiques n'ont pas encore percé tous les petits mystères du fonctionnement de notre système de vision, mais on peut supposer qu'il comporte un certain nombre de niveaux, par exemple comme suit :
-- Pour commencer, des rayons lumineux excitent des capteurs de la rétine.
--- Puis d'autres agents détectent des limites et des textures.
---- Puis d'autres agents encore décrivent des régions, des formes et des contours.
----- Puis certains cadres de souvenirs identifient des objets connus.
------ Ensuite nous identifions des relations structurales entre ces objets.
------- Pour finir, nous établissons un lien entre ces structures et des fonctions et des objectifs.

Il serait donc possible de produire des illusions en agissant à n'importe lequel de ces niveaux. Le plus difficile, ce serait de fabriquer une «image photographique» en activant, de l'intérieur du cerveau, les millions d'agents capteurs de plus bas niveau impliqués dans la vision du monde réel. La méthode la plus simple consisterait peut-être à faire passer exclusivement les agents de vision de plus haut niveau dans les états qui résulteraient de la vue d'une scène donnée (...). Nous appellerons cela un simulus, c'est-à-dire une reproduction des seuls effets de haut niveau d'un stimulus. Avec un simulus aux niveaux les plus hauts, une personne ne se rappellerait peut-être quasiment aucun détail d'un objet ou d'un événement, mais elle pourrait percevoir ses structures et relations les plus importantes et éprouver la sensation de sa présence. Un simulus peut présenter de nombreux avantages sur une image photographique. Non seulement il est plus rapide et plus léger, mais la combinaison de plusieurs simuli permet d'imaginer des choses que nous n'avons jamais vues auparavant, voire d'imaginer des choses qui ne pourraient pas exister."

(Minskii, 1988)

La conjecture dérivant du système de Spinoza, que l'âme ne peut distinguer la présence réelle de sa sensation[20], a pu être montrée de façon expérimentale :

"Depuis les années 1920, l'eidétique n'a plus guère fait parler d'elle. Les observations récentes les plus frappantes proviennent du laboratoire de Wilder Penfield, qui a obtenu ce genre d'images en stimulant électriquement certaines zones des lobes temporaux du cerveau. Telles qu'elles sont décrites par les sujets, les réactions expériencielles, comme les appelle Penfield, s'apparentent à des flashes-back de scènes déjà vécues. Une patiente entendait "chanter un cantique de Noël dans l'église de son pays en Hollande. Elle avait l'impression de se trouver exactement comme elle l'avait été cette nuit de Noël qui remontait à quelques années auparavant". Les sujets s'accordent tous à reconnaître que cette image est plus vivante que tout ce qu'ils parvenaient à se rappeler volontairement : il s'agit non pas d'un souvenir, mais d'un épisode vécu. Tant que l'électrode reste en place, la scène se déroule à sa vitesse naturelle ; la volonté du sujet est impuissante à l'arrêter ni à la faire revenir en arrière. Elle ne s'apparente pas non plus à un rêve ou à une hallucination. Le sujet a parfaitement conscience de se trouver sur une table d'opération et n'est nullement tenté d'adresser la parole aux personnes qu'il "voit". De telles images semblent à peu près aussi achevées que les scènes directement perçues dans le milieu matériel ; à l'instar de ce monde visuel extérieur, elles paraissent avoir le caractère de quelque chose d'objectivement donné, susceptible d'être exploré par la perception active. A cet égard, on peut également les comparer aux images persistantes".

(Arnheim, p. 110-111)

Cet aspect est important car il justifie une approche empirique du monde à partir de ses représentations : il est de toutes façons impossible de distinguer subjectivement un "monde perçu" réel ; donc, étudier les représentations de l'âme, c'est effectivement étudier *tout* ce qui nous est accessible du monde perçu. Les objets que nous cherchons à décrire, qu'on les

[20] "Si le corps humain est affecté d'une modification qui enveloppe la nature d'un corps extérieur, quel qu'il soit, l'âme humaine se représentera ce corps comme existant en fait - ou comme présent pour elle - jusqu'à ce que le corps humain soit affecté d'une autre modification qui exclue l'existence ou la présence du corps en question." (Ethique, livre II, théorème XVII).

appelle idées, représentations ou simulus, semblent bien "exister quelque part" dans les individus.

Nos recherches empiriques confirment, si besoin était, la correspondance entre les associations objectives des objets du monde et les associations subjectives dans les représentations. François Yvon et l'un de nous ont procédé à une approche de description *objective* des consommations alimentaires, calculant les agrégats de produits utilisés par les ménages à partir de leurs fréquences d'achat et de leurs occasions de consommation. Une classification sous contrainte de contiguïté a permis de distinguer comme noyaux "optimaux", sur 150 produits, une vingtaine de "nebulas" ("produits associés dans l'usage concret"), parmi lesquelles on trouve trois associations typiques des petits déjeuners (Lahlou, 1989 ; Yvon, 1990).

> **Nebula 10 :** Pain frais, Beurre, Café moulu normal, Sucre blanc, Confiture, Lait longue conservation, Boisson chocolatée.
> On retrouve ici les produits du petit déjeuner, et plus particulièrement du petit déjeuner "traditionnel".
>
> **Nebula 15 :** Céréales au chocolat, Corn flakes, Muesli, Autres céréales
> Ces produits caractérisent un autre type de petit déjeuner que la nebula 10, mais sont également consommés au goûter.
>
> **Nebula 19 :** Lait en poudre, Biscottes, Pain grillé, Pains spéciaux, Eau minérale grande bouteille, Fruits frais, Eau minérale plate demi bouteille, Margarine allégée, Beurre allégé, Mélange café-chicorée, Café soluble normal, Sucrette, Pain emballé, Moulu décaféiné, Soluble décaféiné.
> Cette nebula regroupe des produits à multiples fonctions : petit déjeuner et repas (pour les pains et les cafés), grignotage et repas (fruits frais, eaux minérales), mais qui tous sont surconsommés par les ménages qui ont des préoccupations diététiques. On notera qu'il ne s'agit pas que de diététique "jeune" (régime de forme) mais aussi de diététique "âgée" (raisons de santé).
>
> (Yvon, 1990)

Une description de la classification obtenue par l'analyse des réponses à la question "petit déjeuner idéal" a déjà été donnée au début de ce rapport et fournit des classes analogues. En voici une autre, obtenue à partir des descriptions des petits déjeuners réels des panélistes.
L'analyse lexicale des descriptions des petits déjeuners (Beaudouin et al, 1993) a permis de retrouver des associations analogues dans la classification des *déclarations* de comportement des 1 000 panélistes[21].
La classe la plus large correspond bien à la nebula 10.

> **Café au lait + pain** (239 individus)
> Cette classe est surtout composée de personnes prenant du café au lait, accompagné de pain, sans beaucoup d'autres produits solides : 78 % déclarent prendre du *café*, 60 % précisent qu'il est *au lait*, personne ne spécifie qu'il est *noir* ; 97 % mangent du *pain*, accompagné de *beurre* ou de *margarine* pour

[21] La correspondance est forcément grossière, en particulier du fait que les nebulas sont des classifications de produits, et qu'un produit ne peut donc se retrouver que dans une seule classe, celle dont il est le plus typique, ce qui n'est pas le cas dans la classification lexicale. On s'attachera donc surtout à examiner les cooccurrences de produits identiques dans les deux méthodes.

78 % et de *confiture* pour 58 %. Dans cette classe très homogène, on ne trouve ni *jus de fruit*, ni *biscotte*, ni *croissant*, ni *biscuit*, ni *gâteau*, ni produit allégé (*beurre allégé, lait demi-écrémé, sucrette*...). Voici quelques réponses caractéristiques de cette classe :
"café au lait et tartines de beurre et de confiture" ;
"café au lait, pain beurre et confiture" ;

Une autre correspond plutôt à la nebula 19, notamment en raison de la présence des produits allégés).

Lait écrémé ou demi écrémé (38 individus)
Elle est essentiellement constituée de personnes ayant déclaré boire du lait écrémé ou demi-écrémé, seul ou bien avec des préparations instantanées (Ricorée, chicorée,...), ou même avec du café. Quant aux aliments solides, ils sont divers mais peu nombreux. Voici quelques réponses caractéristiques :
"lait demi écrémé et café" ;
"lait demi écrémé, café chicorée soluble, pain, margarine, confiture";
"Muesli, lait demi écrémé froid"...

Une autre enfin se rapproche de la nebula 15.

Petit déjeuner anglais (204 individus)
Ce sont les personnes qui prennent des petits déjeuners complets et copieux, constitués surtout de céréales, de laitages et de fruits. Les réponses sont le plus souvent longues et très diverses. La boisson chaude est variée, néanmoins le *thé* est cité par 40 % des individus et *café* par 50 % (le *café* est parfois cité avec une seconde boisson). Le *lait* est aussi cité par 30 %, soit accompagnant une boisson chaude, soit froid avec les céréales. Le chocolat et les boissons instantanées sont peu représentées. Il n'y a pas toujours de boisson chaude, celle-ci étant parfois remplacée par un simple jus de fruit.
Quant aux aliments solides, ils sont plutôt originaux : *croissants, brioches, biscottes et tartines* sont peu nombreux, par contre on trouve beaucoup de *céréales* (28 %), du *fromage* (25.5 %), du *fromage blanc* (14 %), des *yaourts* (33.5 %), des *fruits* (51 %). Voici quelques réponses caractéristiques de cette classe :
"Kellogg's extra ou autres, céréales, jus de fruit, thé aromatisé" ;
"jus de fruit et fruit ou yaourt" ;
"jus de fruit, oeuf à la coque, fromage blanc, café"...

On démontre ici empiriquement un fait assez évident pour chacun : nos représentations sont en rapport avec nos comportements réels. Certains résultats sont cependant moins triviaux. Parmi les nebulas obtenues dans notre enquête de 1988, le classement du chocolat à pâtisser dans la nebula 3, avec les produits festifs, et non pas dans la nebula 12, avec les produits de goûter et de grignotage, nous avait paru un peu étrange, bien qu'explicable facilement (le chocolat à pâtisser est utilisé pour faire de la pâtisserie lors des repas festifs, lesquels sont également caractérisés par les alcools).

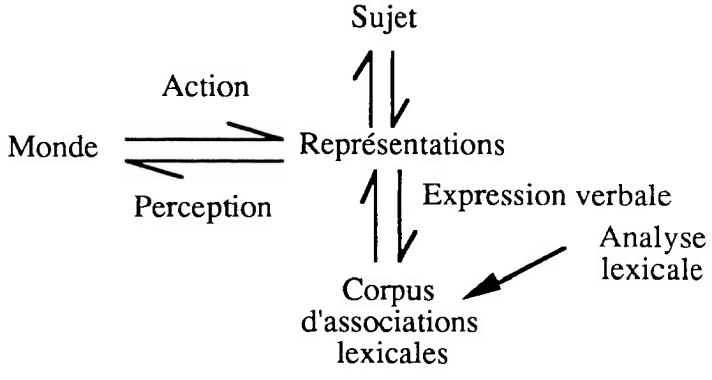
Nebula 3 :
Champagne, Mousseux, Apéritif anisé, Porto, Apéritif, Whisky, Gin, Cognac, Vodka, Chocolat à pâtisser, Vin de qualité
Cette classe rassemble des produits d'apéritifs et de repas soignés. Ils sont doublement proches, puisqu'ils répondent à des fonctions voisines, et que leurs surconsommations sont liées, puisque caractéristiques des ménages qui multiplient les repas festifs.
Les trois classes 1, 2, et 3 s'agrègent ensuite à un niveau supérieur et englobent ainsi tous les produits "festifs". Ceci signifie que si l'on avait coupé l'espace des produits en moins de classes, ces nebula seraient réunies. On aurait donc ici une unique nebula "produits festifs".

Nebula 12 :
Boissons gazeuses sans alcool, Jus de fruits, Autres boissons sans alcool, Chocolat noir, Chocolat blanc, Rocher au chocolat, Pâte à tartiner au chocolat, Gâteaux secs, Biscuit fourré, Barre chocolat, Barre céréale, Bonbon, Chocolat au lait , Chocolat au lait fourré.
Ces produits, utilisés principalement au goûter ou en grignotage, sont les produits surconsommés par les familles, et sont probablement destinés aux enfants. Agrégée ensuite avec la nebula 11, cette nebula rassemble les produits du goûter et du grignotage.

(Yvon, 1990)

Lors d'une présentation de ces résultats, l'un de nos interlocuteurs a remarqué que le fait que le champagne soit un des produits les plus proches du chocolat à pâtisser dans la pratique des usages expliquait sans doute l'impact exceptionnel du spot publicitaire pour le chocolat "Nestlé dessert", où précisément ces deux produits sont étroitement associés dans l'image. Les créatifs de l'agence avaient, guidés par leur seule intuition, mis la main sur une association *qui fonctionnait bien.*

En résumé, à travers les associations lexicales, nous mettons à jour des structures en rapport avec les structures du monde vécu par le sujet. C'est grâce à l'existence des représentations, dont la fonction est précisément de fournir au sujet une image pragmatique du monde sous une forme utilisable et exprimable, que notre méthode d'analyse peut donner des résultats pertinents.



1.3.1. La représentation comme paradigme

Nous considérons la représentation comme un paradigme[22], celui des objets mentaux qui commutent avec le concept central, dans les esprits des membres du groupe social de

[22] "Les unités qui commutent entre elles dans un grand nombre d'énoncés constituent un paradigme. Le phonème français que l'écriture note gn (dans oignon, vigne, etc.) ne figure pas dans le paradigme des consonnes initiales françaises : il ne figure jamais à l'initiale d'un mot français (gnon, gnôle, gniasse, gnangnan, etc., sont des productions populaires qui «tendent» à modifier cette situation, jusqu'ici très marginalement). De même le o de carotte, dit o ouvert (...) ne figure pas dans le paradigme de voyelles finales françaises, mais il a

référence. Comme ce concept central est difficile à cerner, on peut empiriquement définir la représentation comme l'ensemble des idées qui commutent avec le mot qui la désigne. On obtient, en fait, l'expression de ces idées sous forme linguistique. La représentation est alors appréhensible, empiriquement, comme l'ensemble des groupes de mots qui commutent avec le mot central.

On se place donc dans une modélisation de type linguistique, où le modèle préférentiel est bien de type combinatoire (comme le modèle ferroviaire de Minsky) ; mais où la liaison est explicite dans le concept de paradigme, c'est-à-dire qu'elle est empiriquement accessible au lieu d'être fonctionnellement définie. Autrement dit, on définit la représentation de façon topologique : il s'agit d'un réseau d'associations (on sait où elle doit être cherchée, à quoi elle est associée), au lieu de se référer à une description de ses causes ou de ses effets. On peut alors approcher la description des représentations par commutation (en s'inspirant, par exemple, de l'approche des linguistes qui cherchent à isoler des monèmes, ou de l'approche des éthologistes qui cherchent à déterminer des stimuli, en tentant diverses combinaisons de traits pour "voir ce qui marche" c'est-à-dire ce qui évoque une réponse positive[23]. Pratiquement, cela revient à poser des questions du type : et "ça, est-ce que ça correspond bien au concept en question ?" C'est ce que l'on fait, d'une façon assez primaire, en posant des questions à choix multiples.
Nous avons préféré une approche par associations libres, plus ouverte, qui cherche à repérer les trais pertinents en faisant fonctionner le paradigme. Ce fonctionnement repose sur la commutation.

Explicitons. Le paradigme est un réseau d'associations. Donc, si l'on propose un des termes de ce réseau et que l'on demande au sujet de fournir une association, il va fournir un autre terme du réseau. Cet autre terme du réseau va à son tour en évoquer un autre, etc. comme dans l'exemple du cheval de Spinoza. Mais si le sujet a pour consigne de centrer ses associations autour d'un terme, on peut espérer que la commutation va jouer autour de ce terme central, pour faire parcourir au sujet l'ensemble des commutants du paradigme.

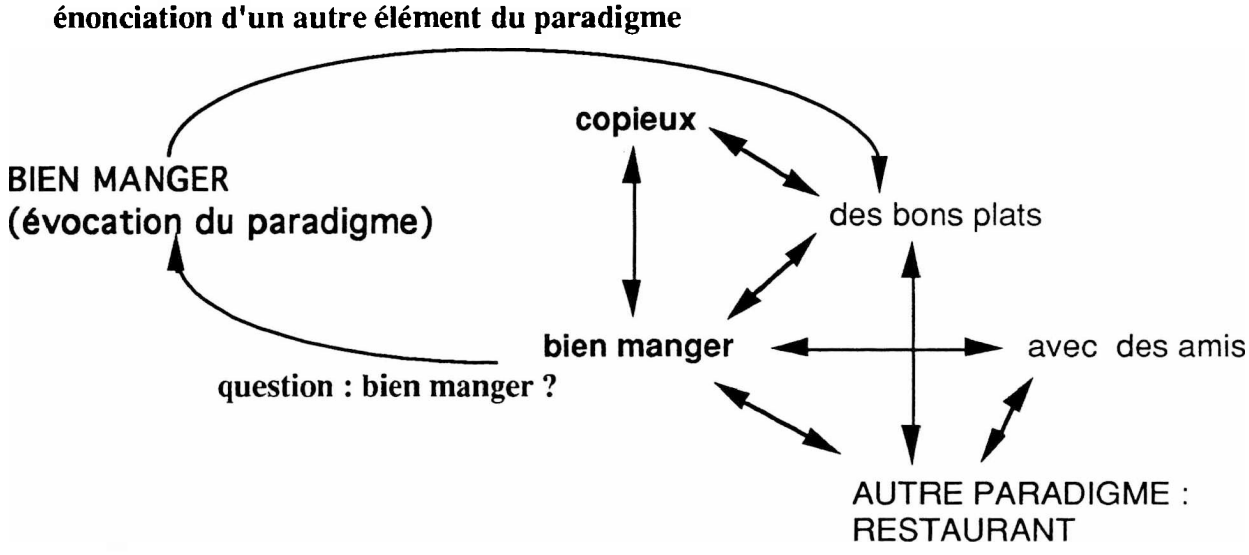
Par exemple, pour cerner la représentation sociale de *bien manger*, on analyse les réponses à une question ouverte : "Qu'est-ce que bien manger ?". Lorsque l'on demande les mots

---

cette propriété structurale en italien : falo, canto, Salo, sont des formes possibles dans cette langue et elles existent". (MOUNIN, 1968)

[23] Nous pensons en particulier aux travaux de LORENZ ou EIBL-EIBESFELDT sur le déclenchement de l'attaque et de la parade sexuelle chez les Cichlidés, dans lesquels on présente aux épinoches mâles des leurres de plus en plus simplifiés, ou aux expériences sur l'évocation du sourire chez le jeune enfant à partir de représentations simplifiées du visage qui montrent les éléments pertinents.

évoqués par ce terme, on obtient ceux qui sont associés de façon proche. Comme on demande plusieurs associations, et que celles-ci s'entraînent l'une l'autre, on peut espérer obtenir tous les commutants psychiques de "bien manger". L'exemple ci-dessous montre de façon schématique comment, si on interroge en associations libres un individu sur un thème, on récupère les autres éléments du paradigme. Un élément du paradigme ("bien manger") évoque la structure, puis celle-ci évoque à son tour un autre élément du paradigme -le même ou un autre-. C'est ce qui est figuré à gauche, en gras.



Mais l'association en chaîne peut très bien amener à évoquer, à partir du premier mot évoqué, un deuxième paradigme, par exemple "avec des amis" évoquera le paradigme de "restaurant"... Il y a risque de dérive associative si les relances sont peu directives.

L'enregistrement de questions ouvertes posées sur le même sujet (Aucouturier et al., 1990) montre de manière assez explicite la façon dont ces évocations se font "en direct". On voit que les interviewés répètent plusieurs fois le mot stimulus, comme pour déclencher un autre mot, embrayer sur une réponse. Prenons deux exemples, où nous soulignons cette répétition de la question, ou *écholalie*, qui semble correspondre chez les interviewés à une sorte de recherche de déclenchement d'associations à partir du stimulus qu'est la question :

Sexe : masculin
Age : 62 ans
Diplôme : Certificat
Profession : Blanchisseur
Taille ménage : 2
Domicile : Banlieue
Q : Si je vous dis "bien manger", quels sont les cinq premiers mots qui vous viennent à l'esprit ?
E : prendre son temps, quoi encore ? **Bien manger** : prendre son temps et puis être en famille. Je sais pas, il y a beaucoup de choses, euh ... ?

Q : Cinq mots !
E : Cinq mots ! (2 s) Je ne vois pas, **bien manger** ? Je ne vois pas, je ... non,
Q : Alors pour vous qu'est ce que "bien manger" ?
E : (1 s) **Bien manger** c'est **manger** euh : (1 s) C'est **manger** euh ... sans gras, sans ..., sans grosses matières grasses bien sûr, sans, sans produits colorants sans ..., **manger** naturel, **manger** des ..., **manger manger** des, naturel oui, **manger** euh, de la verdure, salade, **manger, manger** euh, **manger** la campagne quoi, **manger** moins de conserves possible, voilà
Q : Bien manger pour vous c'est ?
E : Moi ça consiste à ça à mon âge, oui à mon âge, oui euh, j'estime que je dois faire un régime et, j'évite de trop manger, pour moi **bien manger** c'est de manger relativement avec moins de calories possible (1 s) ouais.

Samedi 12 janvier 1991 - 17h - B.H.V.
Sexe : masculin
Age : 60
Diplôme : certificat d'études
Profession : maquettiste en architecture
Taille ménage : 2
Lieu habit. : Paris 3e
Q. Est- ce que vous consommez des produits diététiques , des produits allégés ?
E. Peu, oui et alors ? et ensuite ?
Q. Si je vous dis : bien manger, quelles sont les premières images qui vous viennent à l'esprit ?
E. Euh, allons-y donc, **bien manger**, c'est **manger**, que ce soit des choses parfumées euh légères, des bonnes salades pour commencer euh avec de l'ail par exemple ail, bon pour la viande mais pas forcément tous les jours, du poisson frais et j'évite la boîte de conserve ainsi que les soupes toutes faites. Ca arrive de manière occasionnelle mais enfin c'est des dépannages; voilà et après
Q. Si je vous dis bien manger, quels sont les premiers mots qui vous viennent à l'esprit ?
E. salé, sucré **manger**, se sentir bien comme dirait Harpagon il faut vivre pour manger mais pas, enfin il faut manger pour vivre et non vivre pour manger mais enfin, il est, c'est un bon moment que de passer à table, mais enfin faut pas que ça soit, c'est à l'ensemble de tous les éléments de la vie. Il faut, il faut goûter à tous ces plaisirs mais sans abus quoi, mais se sentir bien. Mon père disait qu'il fallait sortir toujours avec une pointe de faim de la table, pour être bien.
Q. Donc pour vous qu'est-ce que bien manger ?
E. Ca dépend des humeurs ça (4s) d'être avec des gens avec qui je sympathise c'est **bien manger**. C'est toujours la convivialité hein c'est-à-dire se sentir .. soit avec sa compagne, soit de bons amis et passer un bon moment ensemble, c'est ça **bien manger** parce que on peut manger de différentes façons, ça peut être digestif et puis ça peut être (4s) disons, spirituel, mais enfin il faut pas exagérer mais se sentir bien avec les gens c'est ça **bien manger**, oui c'est goûter la vie quoi.

On voit bien sur ces exemples les enchaînements de pensée qui produisent les combinaisons de mots que nous observons finalement. L'intérêt d'utiliser cette catégorie de modèles "par commutation", outre le fait que l'on conserve l'aspect combinatoire, qui intervient au moins à un certain degré, d'après ce que l'on sait de la physiologie cérébrale, c'est que l'on replace forcément la représentation dans le contexte de systèmes articulés pour lesquels on dispose déjà d'un grand nombre de modèles locaux, fondés sur l'observation de phénomènes mentaux, et en particulier le langage. Prises comme combinaisons de mots, les réponses peuvent être insérées dans des processus de calcul mathématique.

1.3.2. L'analyse des énoncés de représentations

L'analyse des représentations comme réseaux d'objets pose la question suivante : comment récupère-t-on une structure en réseau à partir d'un processus d'investigation (ou d'énonciation) séquentiel ? C'est exactement l'inverse de la question posée par Serres au début de son ouvrage "*Le système de Leibniz et ses modèles mathématiques*".

> "Nous ne nous lassons pas de dire que, pour comprendre le système Leibnizien, il convient d'étaler sur une manière d'espace de représentation ce que l'on a coutume d'échelonner le long d'une séquence (p. 31)" (.) "Et puisque l'image que Descartes et la tradition nous ont laissé de leur ordre est celle de la chaîne, qui concrétise pour l'imagination l'unicité de la progression et la liaison des raisons, l'image qui s'impose ici est celle d'un réseau à plusieurs chaînes concourantes, qui présente de multiples entrées et entrecroisements : tapisserie, tissage, broderie ou dentelle (...) Chaque région de ces réseaux est figurée par une sorte de noeud étoilé (de "sommet") dont chaque fil, efférent et (ou) afférent, croise et rejoint tout ou partie de l'ensemble des autres sommets."
>
> (Serres, 1968, p. 14).

Et, pour analyser ce système, Serres doit affronter le problème de la description d'un réseau :

> "Il n'y a pas mille et une manières de comprendre un réseau. Ou l'on circule sur un chemin arbitrairement élu, en conservant, de sommet en sommet, le même fil directeur ; comme disait Descartes, on finit toujours par atteindre l'orée, fût-ce par le plus long. Bien entendu, toute inattention mène au labyrinthe -sans compter que le précepte de suivre une direction constante suppose une référence, un point fixe, c'est-à-dire le problème résolu. De surcroît, parvenu aux lisières, on n'a rien compris des routes de la forêt : un second voyage est une nouvelle aventure.
> Ou bien l'on adopte la technique d'Alexandre à Gordium, on découpe le réseau. Cela signifie que l'on consente d'abord à s'y perdre, à revenir sur ses pas pour *marquer* le chemin. Soit un sommet quelconque du graphe : vers lui convergeront nombre de fils, que l'on accepte de suivre l'un après l'autre, et tous. A les déconnecter alentour du sommet, on obtient une étoile. Tous les points du réseau sont de telles étoiles. soit maintenant un chemin quelconque du graphe ; à le suivre, on rencontre, à terme, un carrefour ; à changer d'orientation en toute rencontre, on revient inévitablement sur ses pas : la route parcourue, découpée à son tour, est un *schéma* polygonal (...). Cela dit, reste à savoir ce qu'est un *point* du dit réseau, étant entendu qu'on vient de s'instruire assez sur ce qu'est un chemin. Voilà notre plan qui n'est autre qu'une carte constituée de ses éléments. Or, dès qu'on dispose d'une carte, le second voyage est fini aussitôt que décidé.
>
> (Serres, 1968, pp. 70-71).

Nous sommes en face du problème suivant. Nous supposons (et nous sommes en cela influencés par une représentation du fonctionnement du cerveau comme réseau de neurones) que les représentations sont des réseaux d'objets mentaux. Or, nous n'avons accès qu'à des descriptions de ce réseau, et ces descriptions sont soumises aux contraintes qu'énonce Michel Serres : elles nous parviennent soit sous la forme de chaînes discursives ordonnées, soit sous forme de descriptions de sommets ou de chemins, c'est-à-dire de morceaux, qui se décrivent les uns par les autres.

Il nous faut donc, à partir de ces parcours, ou de ces morceaux, inférer le tout. C'est comme si l'on avait les morceaux de ce que l'on suppose être un puzzle. Mais la difficulté est que, si pour un puzzle on sait que l'arrangement final a une forme unique connexe, et en général

rectangulaire, dans un plan horizontal, par contre, nous ne savons rien de la forme globale des représentations. Le problème de l'influence de la théorie sur ce que l'on observe joue donc ici à plein. Car on ne trouvera des éléments d'explication que dans la direction où l'on cherche.

1.3.2.1. L'analyse directe des tableaux phrases x mots

Pour être plus clair, supposons que l'objet que nous cherchons à représenter se trouve être "la représentation du beurre".
Il existe différentes formes pour appréhender cette représentation, et chacune nous fournira des réseaux différents.

Dans l'enquête consacrée aux comportements alimentaires de 1988, nous avons demandé à 1 600 personnes de donner leurs associations d'idées sur le beurre. Une question ouverte a été introduite sous la forme :

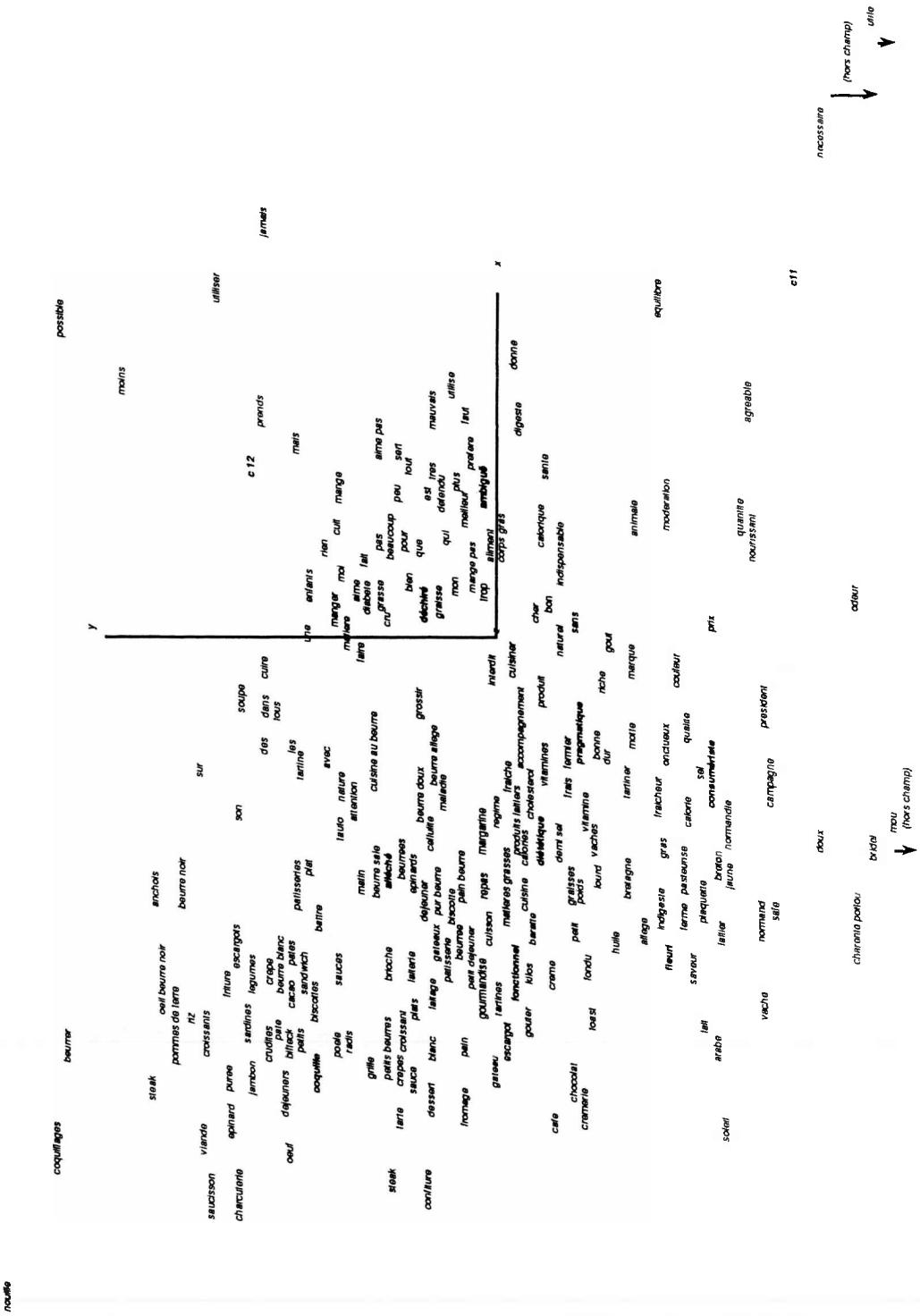
> "Si je vous dis beurre, quels sont les cinq premiers mots qui vous viennent à l'esprit ?"

Nous avons ainsi obtenu 1 600 réponses constituées d'une succession de mots, ou *formes graphiques* (les différentes flexions d'un mot étant distinguées, mais deux mots qui sont employés comme un seul sème -exemple : "matières grasses"- étant comptés comme une forme unique). Nous avons considéré chaque réponse comme une combinaison de formes graphiques. Il devient alors possible de construire un espace ayant autant de dimensions qu'il y a de mots dans le vocabulaire utilisé par l'ensemble des répondants. Dans cet espace où chaque forme graphique représente une dimension, chaque réponse peut être repérée par ses coordonnées obtenues sur chacun des axes par le nombre d'occurrences de cette forme dans la réponse. Cette représentation peut également se traduire par un grand tableau de contingences, avec 1 600 lignes et autant de colonnes que de mots du vocabulaire. Chaque ligne représente une réponse, et chaque colonne une forme graphique. Dans chaque case C(i,j), on met le nombre de formes graphiques "j" dans la réponse "i". Par exemple, la réponse : "tartine, vache, radis, graisse, croissant" sera une ligne avec des zéros partout sauf dans les colonnes correspondant aux formes graphiques *croissant*, *graisse*, *radis*, *vache*, et *tartine* dans lesquelles on aura "1".

Ce tableau de contingence peut être analysé avec les méthodes d'analyse des correspondances multiples, et nous livrer une projection des relations spatiales entre les mots dans un espace ayant un nombre raisonnable de dimensions (deux ou trois) en minimisant la perte d'information. Cet espace est une approximation des champs de connotation du mot beurre, dans le sens où deux mots sont d'autant plus proches dans cet espace qu'ils sont fréquemment

associés dans les réponses à la question initiale. La figure qui suit donne la projection de cet espace dans le premier plan factoriel.[24]

Graphique 15 : Graphe des représentations du beurre (SPAD.T)

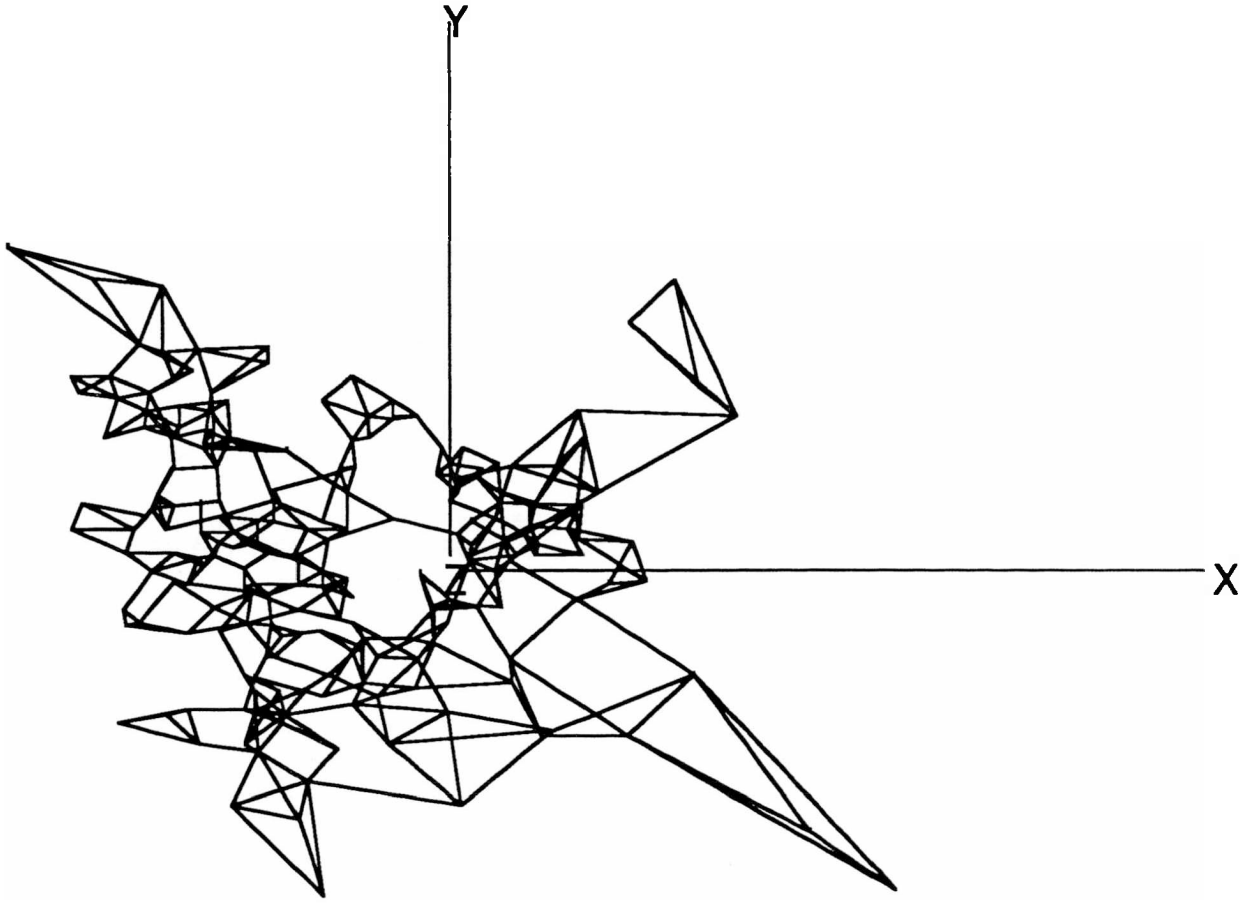


Source : Enquête Comportements alimentaires 1988 - Analyse SPAD.T

24 Voir le rapport de François YVON, *L'analyse lexicale appliquée à des données d'enquête : état des lieux*, Cahier de recherche n°5, décembre 1990. CREDOC.

On a ainsi une première représentation d'un réseau dans lequel est inclus le beurre. Ce réseau est inclus dans un U-langage constitué du lexique français. Afin de donner une représentation plus conforme à ce que l'on imagine être un "réseau", sous forme de points reliés par des lignes, joignons chaque mot à ses trois plus proches voisins[25].

Graphique 16 : Le réseau d'association du beurre

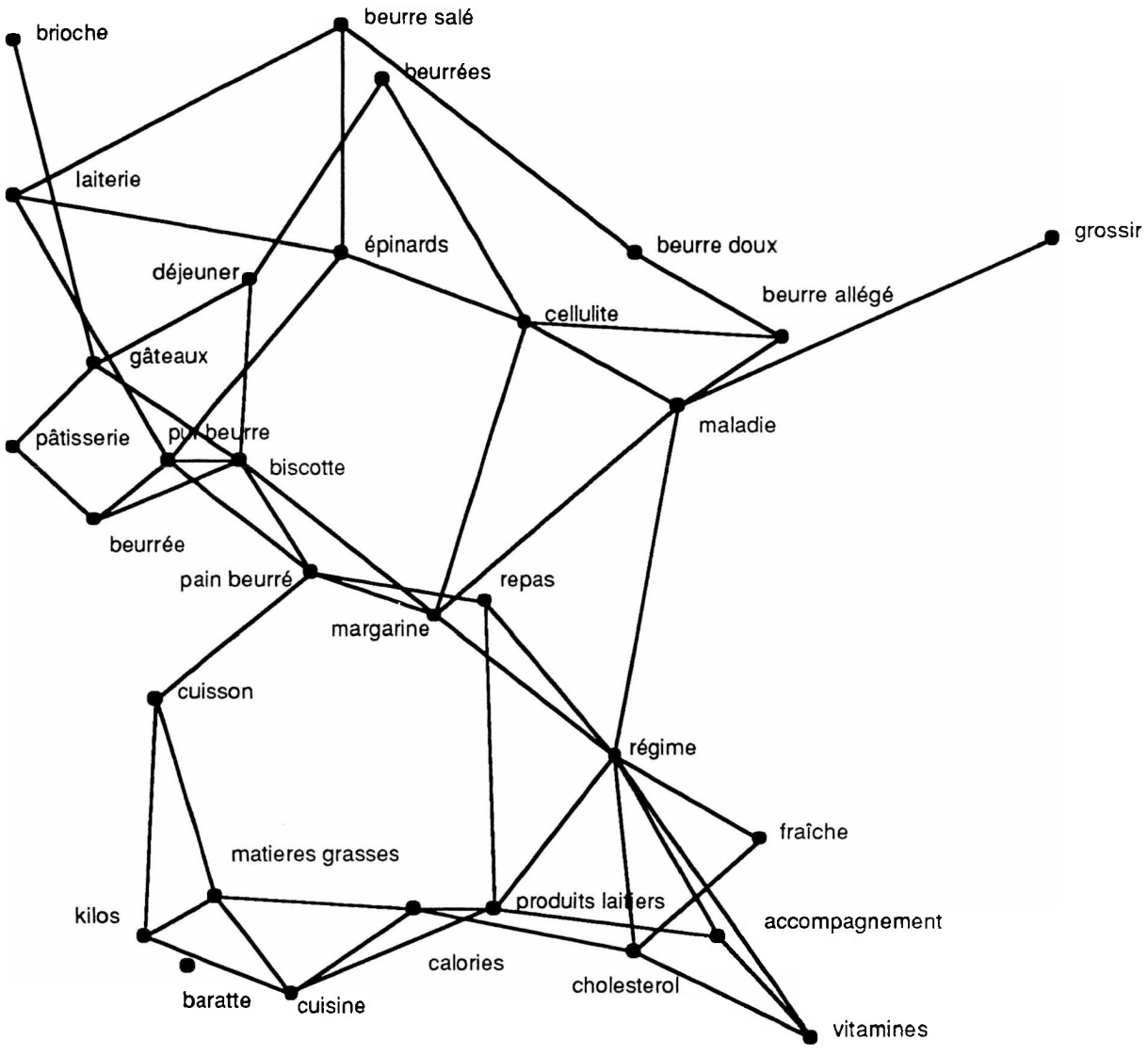


Source : Enquête Comportements alimentaires 1988 - Analyse SPAD.T

Comme ce n'est pas très lisible, agrandissons seulement un extrait.

[25] La distance étant la distance euclidienne classique, calculée dans l'espace dont les dimensions sont les trois premiers axes de l'analyse factorielle, afin de minimiser les effets de parallaxe.

Graphique 17 : Extrait du réseau "beurre"



Source : Enquête Comportements alimentaires 1988 - Analyse SPAD.T

Voilà donc un réseau bien sympathique, sur lequel il est évidemment tentant de construire des interprétations : on remarquera, par exemple, que *gâteaux* et *pâtisserie* ne sont pas loin de "*pur_beurre*", que *régime* et *cholestérol* sont voisins, de même que *matière_grasse*, *calories* et *kilos*, que l'on peut bien construire une chaîne "*grossir-maladie-régime-cholestérol-calories-matières_grasses*", ou encore "*brioche-gâteaux-pâtisserie-beurré-biscotte-pain beurré*", etc. Certes, il existe dans l'espace que nous avons construit sur l'alphabet du lexique français bien des particularités intéressantes. Il faut cependant rester extrêmement prudent lorsque l'on fait des interprétations : on obtiendrait un graphe différent en joignant chaque point aux 5 plus proches (et non pas aux 3 plus proches), et un résultat encore différent si la distance entre les points était calculée autrement, etc.

Quand nous cherchons à repérer la forme de ce réseau, imaginons que nous cherchons à obtenir la structure en parcourant des chemins. Par exemple, nous partons de "beurre", et nous essayons de voir ce qui vient ensuite. Nous pouvons alors considérer les observations obtenues comme un message provenant d'une source.
La théorie de l'information, pour pouvoir étudier le message (et éventuellement chercher à remonter jusqu'au sens de la source) va décomposer le message en éléments, et considérer que l'expression des mots obéit à un processus stochastique, et, si possible, ergodique. L'idée sous-jacente à une telle approche des représentations est que les probabilités transitionnelles traduisent une structure des éléments dans la source, c'est-à-dire qu'elles sont la manifestation soit de règles de génération logique (p[b/a] est déterminée par des théorèmes qui découlent d'une axiomatique), soit d'une topologie (la probabilité transitionnelle de a vers b étant alors d'autant plus forte que a est proche de b), soit d'une structure relationnelle (p[b/a] est d'autant plus forte que b est lié à a).

La question qui se pose est, en admettant que l'on utilise la théorie de l'information pour étudier les messages ordonnés, "que peut-on en inférer sur une éventuelle structure sous-jacente ?". Par exemple, doit-on traduire la représentation sous forme de tableaux de probabilités transitionnelles, ou de lois logiques qui s'appliqueraient à des objets typés par certaines caractéristiques, ou encore par des tableaux relationnels (qui sont un peu un mélange des deux dans la mesure où la relation peut être qualitative) ?

Par ailleurs, vérifier qu'il s'agit de réseaux "déroulés" suppose que l'on a un moyen de savoir comment se "déroule" un réseau. En d'autres termes, dans un modèle combinatoire, où tous les éléments sont liés dans une structure, comment une investigation - qui ne peut pas appréhender la structure d'un seul coup, mais ne peut qu'appréhender les éléments de manière successive -amène-t-elle un élément à la suite d'un autre ? Comment se crée l'ordre temporel à partir d'un ordre relationnel ou géométrique ?

Dans le matériel empirique, nous récupérons un mélange de différents niveaux de langue, or seul le niveau profond (les structures associatives) nous intéresse. Il est évident que, l'expression étant gouvernée par les règles de surface (notamment, syntaxiques), les cooccurrences observées ne reflètent pas exactement la structure des liaisons. L'association *libre*, en éliminant ou diminuant la syntaxe, et en supprimant également une grande part des aspects pragmatiques, permet de se rapprocher de la structure associative profonde.

Une analyse qui porterait directement sur les cooccurrences, ou les corrélations (par exemple l'analyse factorielle sur le tableau initial, comme nous la pratiquions avec SPAD.T) suppose implicitement que les textes obtenus sont une série de signes, *tirés dans un processus ergodique*[26] d'une urne *unique* où les signes sont liés de façon plus ou moins forte. Le tirage est alors considéré comme stochastique (c'est-à-dire qu'il existe un jeu de lois de probabilités qui régit la production des signes), et plus exactement, qu'il s'agit d'un processus de Markov (les cooccurrences obéissent à une unique matrice de probabilités transitionnelles).

On peut en théorie vérifier ces hypothèses (cf. Shannon, 1948, pp. 45-47). Dans la pratique, compte tenu du fait que les phrases sont théoriquement illimitées en longueur, tous les graphes reliant les mots par leurs probabilités transitionnelles estimées sont connexes par construction (il existera toujours une phrase contenant deux mots donnés), et, le nombre de phrases étant infini, le plus petit commun multiple de la longueur des circuits sera effectivement 1, ce qui tendrait à valider l'hypothèse que nous observons un processus ergodique. Cependant, le bon sens nous incline à penser que ces propriétés de nos graphes sont obtenues par construction, et que cette hypothèse ne tient pas dans le cas général, sauf à démontrer que *l'ensemble du langage* est ergodique.

L'hypothèse sur les lois de production et le processus de tirage d'une urne unique est forte. Elle ne nous paraît suffisamment tenable pour bâtir dessus une interprétation sémantique que si l'on est certain de rester constamment à l'intérieur d'un même domaine de représentation, par exemple la représentation d'un objet particulier. En effet, un même mot va voir ses associations changer localement selon les associations pertinentes. Il y a un effet de contexte qui fait que ce même mot va voir ses probabilités d'occurrence changer selon le contexte. Par exemple, *vache* va avoir une probabilité d'apparition très différente dans les contextes *beurre*, *agriculture*, *élevage* ou *pharmacie*. Il en sera donc de même de sa probabilité de cooccurrence avec le terme *brucellose*, simplement parce que ce dernier a *indépendamment* des liens variables avec chacun de ces contextes.

Tant qu'on reste dans un même champ sémantique, on peut supposer que les fréquences de cooccurrence expriment les liaisons relatives entre les mots qui apparaissent dans *ce contexte*. On peut également supposer que ces fréquences relatives dépendent peu de la longueur de la phrase considérée (processus ergodique). Ceci justifie la tentative de construire à partir de ces

[26] Ergodique : qui permet d'extrapoler d'une réalisation unique à la totalité d'un phénomène aléatoire (Le Robert). Plus simplement, dans notre cas : toute phrase est représentative, en termes de fréquences d'occurrence des mots, de la structure profonde (sémantique) du langage.

fréquences de cooccurrences un espace qui sera interprété comme un espace des liaisons (lois de probabilité qui produisent les cooccurrences).
Mais dès qu'on sort de ce cas, il devient probable que plusieurs lois de production ont influencé les cooccurrences observées, chacune des lois étant attribuable à un contexte associatif particulier. On observerait donc, sur des fragments de textes se référant à plus d'un paradigme sémantique, des fréquences correspondant à l'interférence de plusieurs séries de lois de probabilité, dont le nombre nous est inconnu ; ce problème est théoriquement insoluble en raison de l'indétermination du nombre de paramètres.

De cette discussion qui mériterait d'être approfondie par l'examen attentif de graphes de probabilités transitionnelles reliant le vocabulaire, retenons simplement qu'il semble raisonnable d'appliquer la méthode d'exploration statistique fondée sur des matrices de cooccurrence de mots uniquement à des champs lexicaux dont on peut raisonnablement supposer qu'ils ne concernent qu'une représentation unique, avec toutes les approximations que ceci suppose.

L'approche que nous utilisons est différente et ne présuppose pas de loi de production générale du discours.

1.3.2.2. L'analyse par classements de contextes

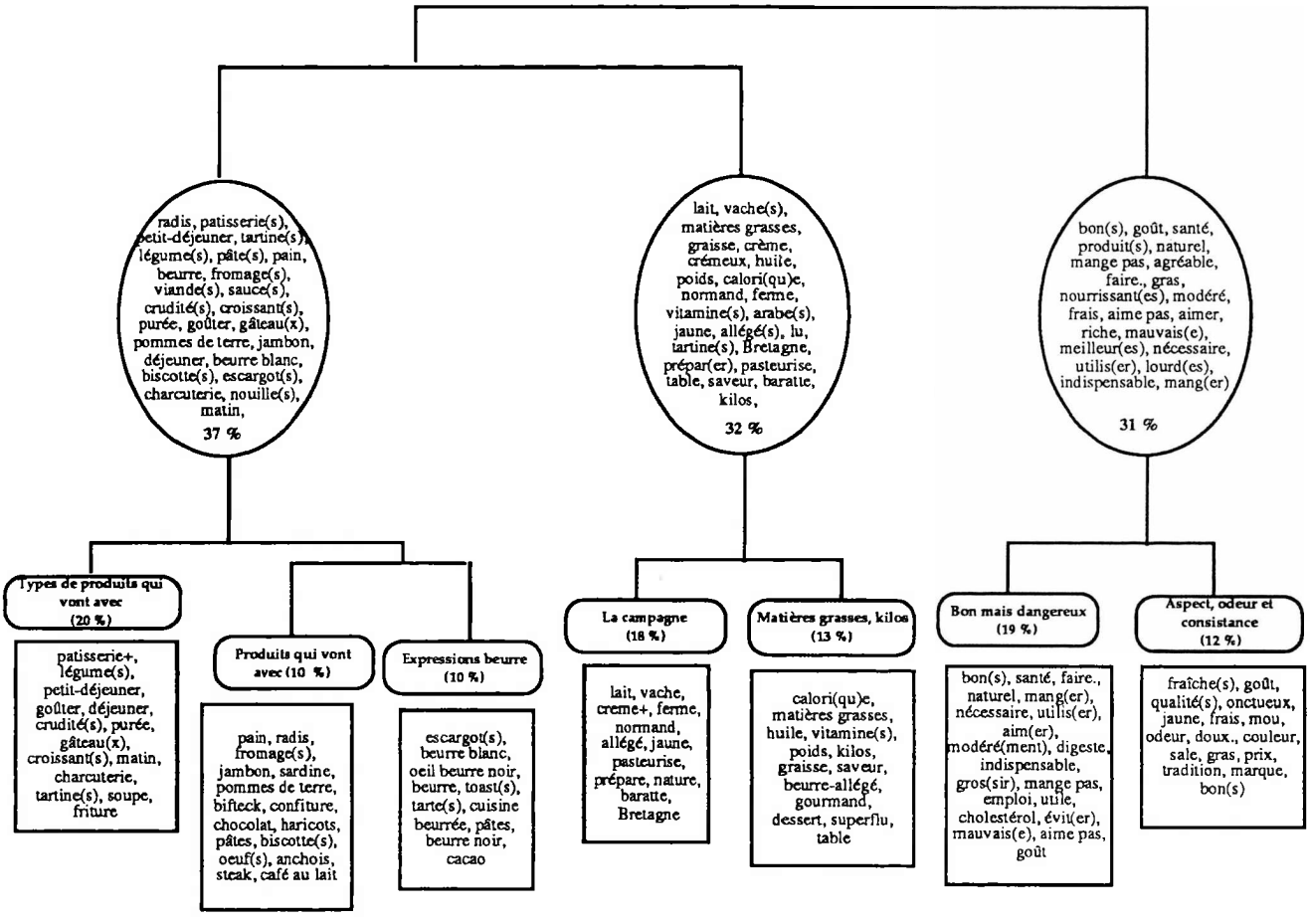
Dans le traitement avec ALCESTE, nous reconstituons, à partir de fragments de trajets sur le graphe général des discours, une matrice de relation, non pas entre les mots, mais entre les unités textuelles. Ceci nous permet de calculer ensuite des distances entre mots, et entre mots et unités. Cette démarche qui servait initialement à déterminer les "mondes lexicaux" correspondant à certaines oeuvres littéraires (Reinert, 1987 a, 1990), est tout à fait adaptée à notre problématique particulière : la détermination des formes de la représentation. Le logiciel est utilisé dans la phase de comparaison analogique grâce à la classification des unités de contexte et dans la phase de détermination des traits saillants du corpus (pour aider à interpréter la classification, ALCESTE propose la liste des termes les plus caractéristiques de chaque classe) (phase 2 et 3a de notre méthode décrite supra).

Sur le plan de l'interprétation sémantique, la méthode consiste à délimiter approximativement, par la classification descendante, des champs relativement homogènes, matérialisés par les classes. Dans ces champs, nous repérons les traits saillants. A partir de ceux-ci, nous

concevons un "paradigme", c'est-à-dire un noyau de sens compréhensible qui contient ces traits[27].

Reprenons avec cette méthode le corpus du beurre, dont l'analyse avec SPAD.T vient d'être présentée dans la section précédente. On obtient une typologie de réponses où chaque classe représente un noyau de sens.

Graphique 18 : Typologie des représentations du beurre (ALCESTE)



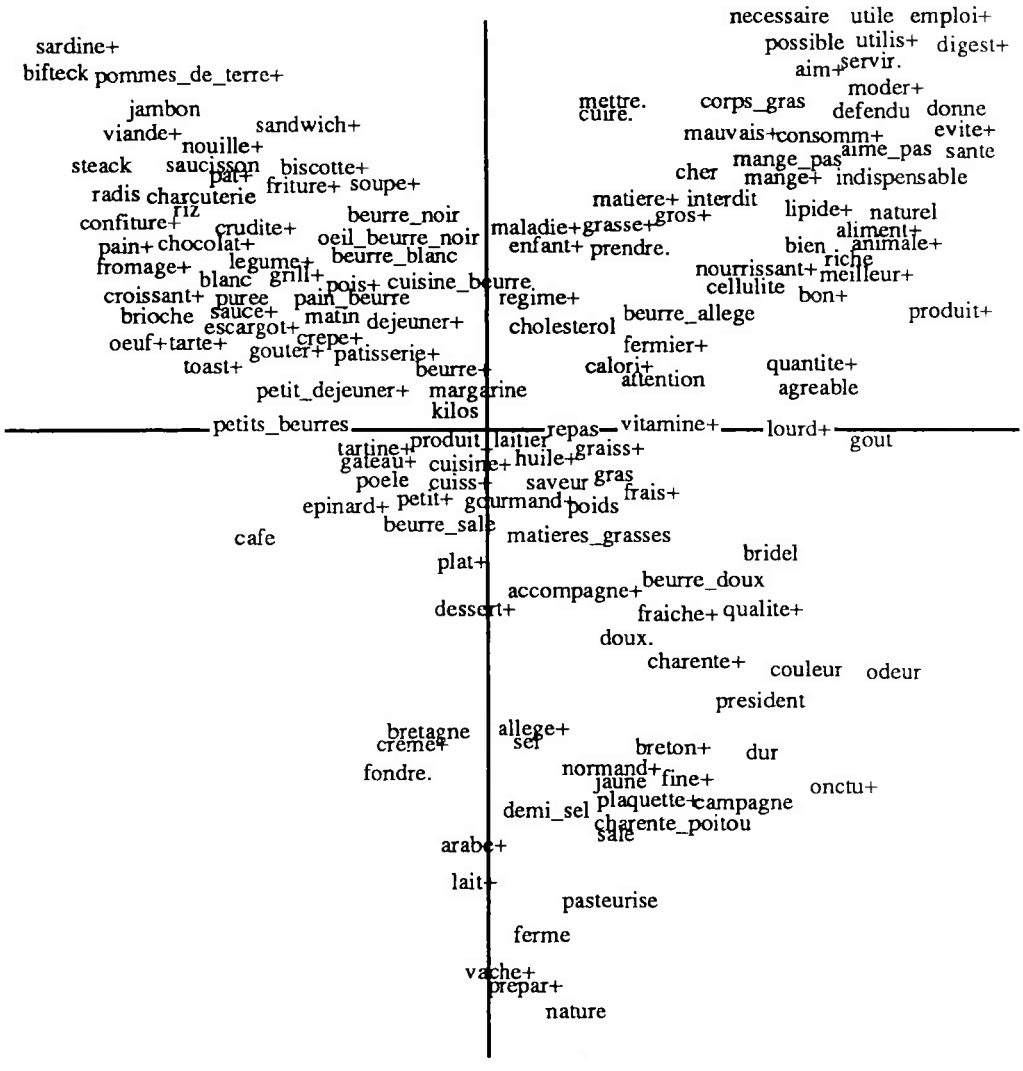
Ensuite, pour voir comment ces paradigmes se situent les uns par rapport aux autres, nous construisons un espace de représentation globale des champs de la façon suivante : nous forçons par une analyse factorielle (lignes = classes, colonnes = mots) une métrique globale qui situe les champs les uns par rapport aux autres, et structure par la même occasion les mots à l'intérieur de chacun de ces champs. Ceci permet d'approximer une structure globale reliant les champs locaux, qui favorise l'interprétation du champ global.

On obtient un graphe qui a une allure sensiblement différente du graphe obtenu avec SPAD.T.

---

[27] Ces étapes correspondent à notre théorie de l'analyse sémantique (cf. supra Troisième partie - 1.2., Construction empirique de catégories par analyse lexicale).

Graphique 18 : Graphe des représentations du beurre (ALCESTE)



Source : Enquête Comportements alimentaire 1988 - Analyse ALCESTE

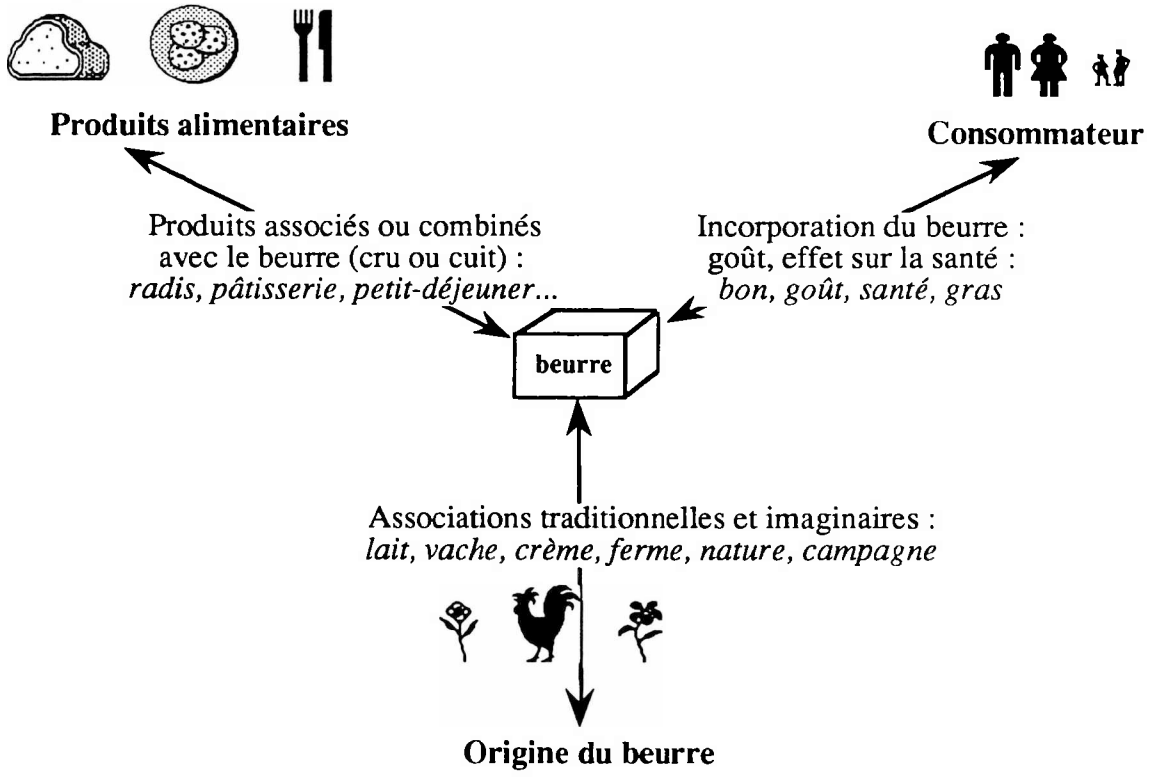
On arrive plus facilement à identifier les grandes dimensions de la représentation du beurre et ceci grâce au type de construction utilisée. En effet, l'analyse factorielle est effectuée après la classification sur une matrice où l'information est déjà synthétisée (matrice qui contient en ligne les mots et en colonne les classes des réponses). Les deux premiers axes de l'analyse factorielle (sur lesquels sont projetés les mots dans le graphique précédent) représentent plus de 50 % de l'inertie : les risques d'erreurs dans l'interprétation des proximités sont beaucoup plus faibles que sous SPAD.T où l'inertie des deux premiers axes est souvent inférieure à 10 %.

D'autre part, la lecture du graphe est rendue plus lisible par le fait que seuls apparaissent les termes qui ont un coefficient d'association aux classes de la typologie assez élevé. Le graphique ne représente que les traits des paradigmes les plus significatifs. Il est ainsi plus

facile de reconstruire les noyaux de sens. Le graphique ci-dessous propose une synthèse du graphe global, montrant les trois paradigmes associés au beurre :

- le beurre est une substance combinée avec d'autres aliments au niveau de la préparation (recettes) ou de la consommation (beurre-coquille) ;
- le beurre lorsqu'il est consommé a des effets immédiats (perception de la saveur, du goût) et des effets secondaires (poids, santé) négatifs ou positifs ;
- la chaîne associée au beurre est une chaîne de remontée aux origines du produit. Le beurre renvoie à une représentation de l'univers rural.

Graphique 20 : Les trois dimensions de la représentation du beurre



La représentation est le paradigme des objets mentaux qui commutent avec le concept. C'est un réseau d'associations. Comme tout objet, les représentations peuvent être considérées comme un tout ou comme une combinaison de parties. Selon ce point de vue, on analysera la représentation comme un réseau unique ou bien comme la juxtaposition de plusieurs réseaux. C'est la différence majeure entre SPAD.T et ALCESTE. SPAD.T considère l'ensemble des réponses comme un tout lorsqu'il effectue une analyse factorielle et cherche à positionner tous les termes les uns par rapport aux autres. Pour ALCESTE, au contraire, l'objectif est d'identifier dès le départ les réseaux d'association locaux. Ensuite, les différents paradigmes sont positionnés les uns par rapport aux autres.

ALCESTE est donc plus apte que SPAD.T à identifier les différents sous-paradigmes constitutifs de la représentation. SPAD.T aboutit aussi à des typologies (en effet, sur la base de l'analyse factorielle, une classification permet de mettre en évidence des sous-paradigmes) mais elles sont plus fragiles dans la mesure où elle sont construites sur l'analyse factorielle qui considère les réponses comme un tout. Dans l'analyse sur le beurre de SPAD.T (Yvon, 1990), certaines classes correspondent aux classes obtenues par ALCESTE (les classes de produits alimentaires, l'évocation de la campagne) mais globalement le contenu des classes est moins homogène. Il est plus difficile d'inférer des noyaux de sens et donc d'interpréter les classes. ALCESTE est mieux adapté que SPAD.T à la démarche interprétative proposée ci-dessus.

2. L'EXEMPLE ET LE CONCEPT : PROTOTYPES OU DESCRIPTEURS

En confrontant de nombreuses analyses et en appliquant (encore intuitivement) la méthode d'interprétation proposée ci-dessus, nous avons pu observer que dans les descriptions d'objets par enquête, il existe deux types de paradigmes qui correspondent à deux modes de pensée distincts.

Les questions fermées ont en général des objectifs purement opératoires : connaître l'opinion des électeurs, consommateurs, etc. pour effectuer des prévisions et s'adapter. Les questions ouvertes permettent d'accéder aux représentations mentales et de comprendre les processus qui entrent en jeu pour appréhender le réel par la pensée et par la parole. Les questions ouvertes nous donnent autant d'informations sur ce qui est vu que sur la manière de voir.

Les nombreuses analyses de réponses à une question ouverte que nous avons faites nous ont conduits à émettre l'hypothèse suivante : il semblerait que deux types de processus cognitifs soient mis en jeu pour répondre à une question :
- une approche prototypique : la personne interrogée énumère des noms d'objets ;
- une approche conceptuelle : l'enquêté a recours à des descripteurs (ou critères) pour répondre.

De plus, nous avons pu dégager certaines régularités dans la démarche adoptée par les enquêtés pour répondre à une question : quand on répond en terme de prototypes à une question on a tendance à répondre de la même manière à toutes les autres. Quels facteurs influencent le choix d'une démarche particulière ?

Avec l'enquête "Qualité mesurée, qualité perçue" (Beaudouin, Hébel et Lahlou, 1992) effectuée dans le cadre d'un programme de recherche en design du Ministère de la Recherche et de l'Espace, nous disposions d'un matériau empirique remarquable pour vérifier notre hypothèse et trouver des facteurs d'influence. En effet, l'enquête comportait six questions ouvertes dont la formulation était similaire :

Si je vous dis … ["produit de bonne qualité" / "produit de mauvaise qualité" / "fromage de qualité" / "voiture de qualité (ou idéale)" / "habitat de qualité (ou idéal" / "banque de qualité (ou idéale)"]

quels sont les premiers mots qui vous viennent à l'esprit ?

Ces questions ont été posées à un échantillon de 900 personnes représentatives de la population française. Pour la voiture, l'habitat et la banque, on a remplacé sur la moitié de l'échantillon le syntagme adjectif *de qualité* par l'adjectif *idéal*, dans le but d'enrichir le corpus et de mieux comprendre les écarts entre la qualité et l'idéal dans l'esprit des consommateurs.
Nous avons donc pu, sur un nombre important d'exemples, confirmer des intuitions nées lors de nos premières analyses lexicales.

L'analyse des réponses aux six questions ouvertes a confirmé qu'il y avait bien deux manières dominantes distinctes de répondre à des questions portant sur la représentation mentale d'une catégorie d'objets :
- soit en donnant une liste de critères ou de descripteurs ;
- soit en proposant des objets du monde qui sont, aux yeux des enquêtés des exemples de produits de qualité ou des prototypes.

Dorénavant, nous utiliserons les termes de descripteurs et de prototypes mais il nous faut préciser le sens qu'ils acquièrent dans ce contexte particulier. Le prototype est d'après le dictionnaire "un type, modèle ou exemple parfait". Le prototype est une sorte de construction mentale, le meilleur représentant possible d'un concept.

> Quand on considère le règne animal, et qu'on s'aperçoit que, parmi les quadrupèdes, il n'y en a pas un qui n'ait les fonctions et les parties, surtout intérieures, entièrement semblables à un autre quadrupède, ne croirait-on pas volontiers qu'il n'y a jamais eu qu'un premier animal, prototype de tous les animaux, dont la nature n'a fait qu'allonger, raccourcir, transformer, multiplier, oblitérer certains organes ?
>
> DIDEROT, De l'interprétation de la nature, XII.

Actuellement, il existe deux acceptions possibles du terme prototype : la première, qui est celle que nous allons adopter, est "représentation du meilleur exemplaire d'un concept"[28]. La

[28] FORNEL (Michel de), "Sémantique du prototype et analyse de conversation", *Cahiers de linguistique française*, n° 11, Genève, 1990.

seconde moins adaptée à notre propos considère que le prototype est un "schéma cognitif, une représentation des propriétés courantes associées aux réalisations d'un concept"[29].
Pour nous le prototype n'est pas le meilleur exemplaire universellement reconnu, mais le meilleur exemplaire retenu par l'individu. En donnant des exemples, l'enquêté propose les objets qui à ses yeux incarnent le mieux le concept : des prototypes individuels.

Bien sûr certaines réponses sont une combinaison de ces deux manières de répondre (par descripteur ou par prototype), comme on le voit sur les exemples suivants :

La Poste, bon accueil, gentils (banque)
B.N.P., bien accueilli quand on arrive et pas de problème (banque)
Nourriture, *vin*, de bonne fabrication, solidité, fiabilité, rapport qualité-prix (produit de bonne qualité)
Prix, qualité, *poulet fermier* (produit de bonne qualité)

Ainsi dans une réponse, il peut y avoir des éléments relevant des deux types d'approche. L'enquêté qui s'engage dans un type de réponse peut bifurquer vers un autre. Il serait d'ailleurs intéressant d'étudier ces déviations cognitives.

Mais là n'est pas notre propos, étant donné que dans la majorité des cas on peut dire qu'une réponse est soit prototypique soit descriptive. La ligne de partage entre ces deux types de réponses est similaire à celles qui oppose les définitions en extension aux définitions en compréhension ou en intension. Revenons aux définitions que le Grand Robert donne de ces concepts :

> **Extension** : (Déb. XXe). Log. Ensemble des objets concrets ou abstraits auxquels s'applique un concept (ensemble des objets qu'il peut désigner, dont il est l'attribut), une proposition (ensemble des cas où elle est vraie ; par suite, ensemble des hypothèses dont elle peut être la conséquence), une relation (ensemble des systèmes de valeurs attribuées aux termes généraux qui vérifient ces valeurs), d'après Lalande.
>
> Le Robert Électronique

> **Intension ou compréhension** : Log. Ensemble des caractères qui appartiennent à un concept, le constituent et servent à le définir. Définition par intension.
>
> Le Robert Électronique

Les réponses prototypiques cherchent à donner une liste d'objets (dans notre cas des produits) auxquels on peut attribuer le concept de qualité ou d'idéal. Ces réponses par l'exemple manifestent un ancrage dans le monde de référence important : un produit de qualité est

[29] idem.

forcément un produit dont l'existence est attestée puisque son nom est donné par l'enquêté. On peut distinguer divers degrés de réalité dans les réponses données. Les réponses :

B.N.P., Société Générale

renvoient à des institutions financières encore clairement délimitées, tandis que :

roquefort, camembert

étant donné le nombre de variétés distinctes sur le marché, sont des objets déjà plus abstraits. On retrouve ici la problématique classique du prototype en psychologie cognitive. L'objet de référence peut être une catégorie plus ou moins large.

Les réponses en terme de critères donnent des listes de caractéristiques propres à l'objet ou au concept. Elles manifestent un degré d'abstraction supplémentaire dans la mesure où elles impliquent de se détourner des objets du monde pour mettre en évidence les ressemblances, les caractéristiques propres à tous les produits qui méritent de recevoir l'appellation "de qualité" pour l'enquêté.
Cette alternative prototype / descripteur se retrouve dans la plupart des questions que nous avons posées sur des objets. Ainsi, sur le *bien-manger*, une classe de description pragmatique des repas en terme de produits (entrée, plat, dessert) s'oppose à des classes de description plus conceptuelles (Lahlou 1992).

2.1. Personnes âgées et prototypes

Dans nos enquêtes, les réponses prototypiques sont caractéristiques des personnes de plus de soixante ans. Plus les personnes sont âgées, plus elles ont tendance à s'exprimer de cette façon plutôt que d'avoir recours à des critères pour définir la qualité.
D'autre part, on remarque aussi que moins les personnes sont diplômées, plus elles ont tendance à répondre à l'aide de prototypes.

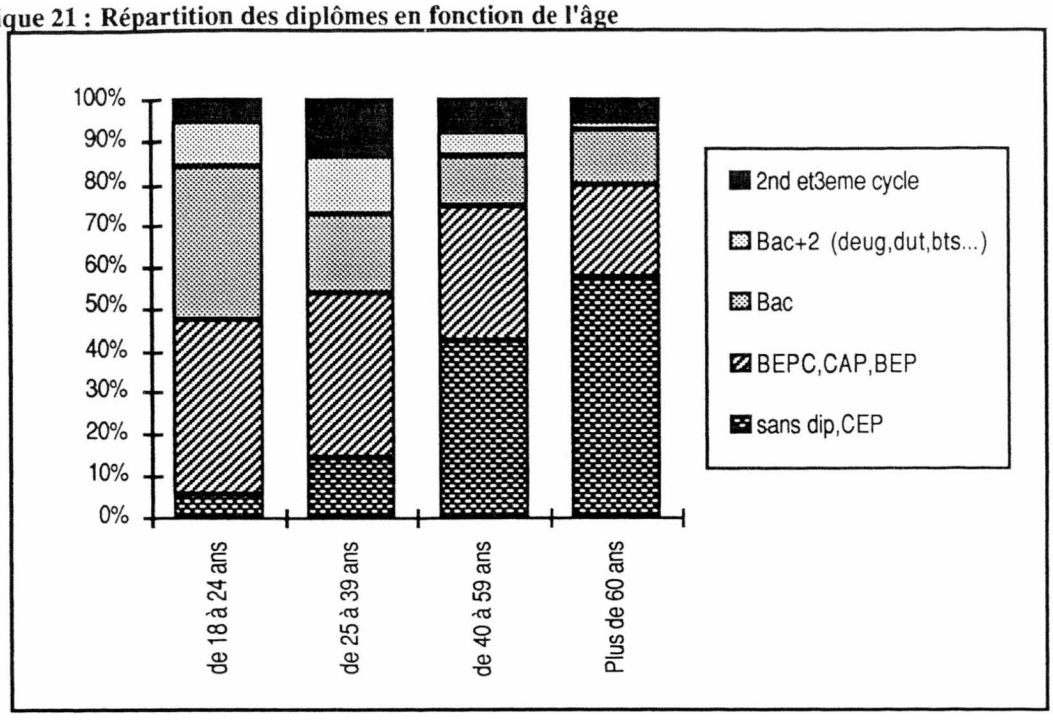
On peut à ce propos émettre deux hypothèses :

• L'expérience permet aux personnes âgées de mieux connaître leurs attentes et les produits qui répondent le mieux à ces attentes.
Dans un souci d'efficacité pragmatique, elles auront tendance à répondre en donnant la liste des produits qui satisfont leur demande, c'est-à-dire la liste des produits qu'elles choisissent pour elles-mêmes ou la liste de ceux qu'elles conseilleraient à leur interlocuteur.

Les prototypes, dans la mesure où ils sont connus du locuteur et de l'interlocuteur sont des outils de communication infiniment plus efficaces que les critères. Même s'il existe une grande variété de camemberts, répondre "camembert" à la question "Pour vous, qu'est-ce qu'un fromage de qualité ?" est plus précis que de répondre "un fromage coulant à point...", la représentation de l'objet est plus immédiate, parce qu'elle est sensoriellement complète (elle est rattachable à une expérience directe et totale). Tandis que la description par critère implique de faire une abstraction dans le tissu de l'expérience pour ne retenir qu'une modalité.

• L'esprit humain fonctionnerait plutôt à l'aide de prototypes que de descripteurs, car ce sont des objets plus faciles à manier. L'école socratique avait explicité cette préférence de l'esprit pour l'exemple et les images. Même pour expliciter leurs concepts, Socrate et ses disciples recourraient fréquemment à la métaphore comme figure la plus apte à permettre le passage du sensible à l'intelligible. Toute réponse conceptualisée est la marque d'un effort de l'esprit pour se déprendre de la matérialité et se tourner vers l'idée.
Cet effort d'abstraction nécessite un apprentissage qui, dans nos sociétés, est dans une large mesure pris en charge par l'éducation scolaire. Or depuis l'après-guerre, le niveau moyen d'éducation n'a cessé d'augmenter grâce à la généralisation de la scolarisation (école obligatoire jusqu'à 16 ans). En conséquence, l'écart entre le niveau moyen d'éducation des jeunes générations et des anciennes s'est accentué. Sur l'échantillon de 900 personnes retenu pour l'enquête sur la qualité, il est évident que plus l'âge augmente, plus la proportion de non diplômés augmente. En effet, près de 60 % des personnes de plus de 60 ans n'ont aucun diplôme ou un CEP.

Graphique 21 : Répartition des diplômes en fonction de l'âge



Source : Enquête Qualité

La façon de répondre à une question ouverte témoignerait dans une certaine mesure d'un niveau de diplôme.
Finalement, eu égard à cette seconde hypothèse, le fait de répondre en termes de noms de produits serait plutôt lié au niveau de diplôme qu'à l'âge. Ce ne serait donc pas une conséquence du vieillissement, mais un effet de génération qui reflète le niveau d'instruction.

Les deux hypothèses émises précédemment montrent que pour deux raisons distinctes, ce sont les personnes âgées qui s'expriment de manière prototypique. Finalement ces deux interprétations ne sont pas contradictoires. Le recours à des réponses prototypiques manifeste une certaine distance face à la pensée abstraite (plus grande chez les personnes âgées car moins diplômées) mais aussi un souci d'efficacité dans l'acte de communication (plus fréquente chez les personnes âgées moins soucieuses de leur image "sociale").

Puisque de toutes manières en dépit de la diffusion de l'enseignement, il semble plus facile de répondre à l'aide de prototypes que par critères, comment se fait-il qu'en dehors des personnes âgées la proportion de réponses plus conceptuelles soit la plus forte ?
Qu'est-ce qui peut pousser les consommateurs à s'exprimer en termes de descripteurs alors que les prototypes sont tellement plus faciles d'usage ? On imagine que dans des domaines où l'existant est très pauvre, il serait plus difficile de répondre en terme de prototypes, mais sur

les marchés que nous avons choisis, l'abondance de l'existant aurait dû entraîner un nombre plus grand de réponses en terme de prototypes.

On peut émettre l'hypothèse qu'il est plus valorisant pour l'individu de répondre à l'aide de descripteurs, qu'à l'aide de noms de produits. L'éducation a pour tâche de donner plus de valeur au concept qu'à l'image : l'exemple est second, il ne vient que pour illustrer la pensée. Ainsi, la personne enquêtée se ressent dans une situation qu'elle compare au cadre scolaire et aura tendance à répondre en fonction de l'enquêteur, qui, à ses yeux, détient une légitimité proche de la légitimité scolaire. C'est dans un souci de présentation de soi que les personnes répondent en terme de critères.
Les réponses en termes de descripteurs marquent un souci de présentation de soi et manifestent un désir de distinction. Les personnes âgées ne feraient pas cet effort, parce que leur souci de positionnement social serait moins vif que pour les générations plus jeunes.

2.2. Idéal et prototypes

Nous avons pu observer que les réponses en terme de noms d'objet ou de prototypes étaient plus fréquentes quand la formulation de la question était du type "produit idéal". Inversement, à la question "produit de qualité", une majorité de personnes répondent à l'aide de critères, concepts ou adjectifs qualificatifs. Ce résultat est contre-intuitif, puisque, a priori, l'objet idéal n'existe pas et devrait appeler une réponse plus conceptuelle. Mais il faut considérer le mécanisme de réponse à la question qui produit les réponses attestées, et non pas spéculer sur le contenu théorique de la question.

Nous l'avons vu, les enquêtés ont tendance à s'ancrer sur la question pour déclencher leur réponse, et les mots clefs de la question vont donc engendrer des réponses issues de l'ensemble de leur champ de connotations et, en particulier, les connotations vécues.

Il semblerait que le mot *qualité* soit inconsciemment perçu comme une catégorie de pensée, héritage d'une longue tradition philosophique (c'est le premier sens de *qualité* dans le Robert) :

> *I. Philos. La qualité : une des catégories* fondamentales de l'être ; manière d'être, attribut propre de l'être (par oppos. à l'être, à la relation, etc.).
> Le Robert Électronique

Par conséquent, le sujet répond en s'efforçant de rester dans le même registre, c'est-à-dire en ayant recours au concept et aux critères.

Par ailleurs, il existe un certain nombre de procédures ou d'a priori pour déterminer la qualité. Ceux-ci prennent la forme de critères largement diffusés par les fabricants dans leur communication publicitaire. Il existe donc un réservoir potentiel de critères dans les champs de connotation du mot qualité.

Inversement, l'adjectif *idéal,* qui appelle peu de critères normatifs, semble être plutôt associé dans l'esprit de l'enquêté à une catégorie de l'imagination qui autorise l'expression de la subjectivité. Ainsi, dans ce cadre, une réponse en terme d'image, de représentation est légitime aux yeux des enquêtés. Les réponses à la formulation "idéal" sont plutôt des images visuelles, concrètes plutôt qu'une liste de critères abstraits.

Habitat idéal :

une maison avec un grand jardin, une piscine et des fleurs

Habitat de qualité :

confort, espace, près du centre-ville

Finalement, la formulation idéale permet de recueillir une information beaucoup plus variée dans la mesure où le terme *idéal* est moins marqué que celui de qualité. On pouvait donc imaginer que la formulation idéale entraînerait des débordements d'imagination. De fait, l'idéal n'est vécu que comme une légère extrapolation du réel ; la référence est toujours le monde tel qu'il est accessible à l'enquêté.

En résumé, pour l'habitat, la voiture, le fromage, (c'est un peu moins vrai pour les banques, peut-être parce qu'elles ont moins d'impact sur l'imaginaire), les réponses en terme de prototypes sont beaucoup plus fréquentes quand la question porte sur l'idéal, les réponses en terme de critères quand la question porte sur la qualité (Beaudouin, Hébel et Lahlou, 1992)).

L'analyse des questions ouvertes portant sur des objets nous a permis de mettre en évidence que l'ensemble des réponses s'organisaient autour de deux types de paradigmes (qui semblent refléter deux processus cognitifs distincts). Le premier plutôt conceptuel s'organise autour de traits-critères, le second autour de traits-noms d'objets.

CONCLUSION

Beaucoup reste à faire. On peut néanmoins tirer de ces expériences quelques conclusions.
Sur le plan théorique, l'analyse lexicale par contexte semble bien appropriée à l'exploration des représentations mentales et sociales. Elle correspond à l'application de la méthode d'investigation scientifique à des objets modélisés comme des combinaisons de traits. Le principe heuristique sous-jacent est l'abduction de catégories sémantiques à partir de traits lexicaux statistiquement saillants.
Les lois de production correspondantes sont l'expression linguistiquement correcte des éléments du paradigme qu'est chaque représentation.
Une attention particulière doit donc être portée aux mécanismes de production linguistique (en situation d'enquête, notamment). Nos premières investigations montrent qu'il existe au moins deux grands types d'approche cognitive qui régissent cette expression : une approche prototypique (extensive) et une approche par critères (compréhensive).

Sur le plan méthodologique, on retiendra notamment qu'il est souhaitable de prendre une certaine liberté vis-à-vis du corpus, afin de traiter informatiquement les seuls traits pertinents pour la problématique considérée et limiter ainsi le "bruit" informationnel contenu dans l'expression lexicale. La lemmatisation, la réduction sémantique, la projection dans un sous-langage idoine sont les approches que nous avons explorées ; elles s'avèrent utiles.

Sur le plan technique, l'utilisation de lemmatiseurs sophistiqués, bien qu'elle n'apporte pas une amélioration sensible des résultats, nous semble souhaitable, d'autant qu'elle est maintenant parfaitement opérationnelle. Les réseaux sémantiques sont aussi une voie à explorer. Dans ce domaine, les progrès sont rapides mais les applications de la méthode restent à l'heure actuelle encore problématiques.
Dans les méthodes de recueil, il est manifeste que des biais importants sont apportés par les modes de questionnement. Ce point doit être pris en compte dans les choix de terrain.

Enfin, les perspectives d'application de la méthode d'analyse par contexte à tout type de description par traits, en particulier les enquêtes "fermées" classiques, méritent d'être explorées.

BIBLIOGRAPHIE

ANCORI, Bernard.(1983).- "Communication, information et pouvoir". *Information et communication*. Séminaires interdisciplinaires du Collège de France. Sous la direction de A. Lichnerowicz, F. Perroux et G. Gadoffre. Paris, Maloine.- pp. 59-84.

ANDREW, R. (1963).- "The origin and evolution of the calls and facial expressions of the primates", *Behaviour*, 20.- pp.1-109.

ANSART, P. (1987).- "Le concept de représentation en sociologie", L. Marbeau, F. Audigier (eds.), *Seconde rencontre sur la didactique de l'histoire et de la géographie*. (Table ronde sur le concept de représentation dans différentes sciences de l'homme et de la société). Paris, Presses Universitaires de France, INRP.

ARNHEIM, Rudolf (1976).- *La pensée visuelle*.- Flammarion.- p. 37.

AUCOUTURIER, Anne-Lise, BEAUDOUIN, Valérie, BLOT, Isabelle, FAIVRE, Didier, LAHLOU, Saadi, MICHEAU, Julie (1991).- *Nature et traitement statistique des données textuelles. Réflexions méthodologiques* -. Paris, CREDOC, Cahier de recherche, n°24, Département Prospective de la Consommation.

BATESON, G. et JACKSON, D. D. (1964).- "Some varieties of pathogenic organizations" (Rioch, David Mc K. (ed), Disorders of communication), *Research Publications A. R. N. M. D.*, 42.-pp. 270-283.

BEAUDOUIN, Valérie, HEBEL, Pascale, LAHLOU, Saadi, LE BIHAN, Hervé.- "Enquête sur la qualité perçue", CREDOC (Dépt. Prospective de la Consommation), Les Ateliers-ENSCI, MIND MOVERS (1992).- *Comprendre et évaluer la qualité* -. Paris, CREDOC, Cahier de recherche, n°39.

BEAUDOUIN, Valérie (CREDOC), GILLES Marie-Odile, LAHLOU, Saadi (CREDOC), MARSAL, Franck, PEYRE, Laurence, WATTRELOT, Stéphanie, YVON, François (ENST) (juin 1993).- *Analyse automatique du discours : le petit déjeuner*. Rapport de groupe de travail de 3e année à l'ENSAE.

BEAUDOUIN, Valérie (1993).- *Analyse lexicale et stylistique : Gravitations de Jules Supervielle* - Paris, CREDOC, Cahier de recherche, n°49.

BEDECARRAX, Chantal (1988).- "Quadri-décomposition : modèle général". - Etude du Centre Scientifique IBM-France, F 0132, Paris.

BENZÉCRI, Jean-Pierre, LEBART, Ludovic, REINERT, Max (1981).- *Pratique de l'analyse des données, Linguistique et lexicologie*. - Paris, Dunod.

BERHAKIS, Théodoros, NICOLAIDIS, Efthymios (1990).- "Typologie des prologues des livres grecs de sciences édités de 1730 à 1820" *Les cahiers de l'analyse des données*, vol XIV, n° 1.

BETBEZE, Jean-Paul, LAHLOU, Saadi, MAFFRE, Joëlle (1987).- Les produits alimentaires à l'horizon 1995 : du surgelé au «nouveau frais». Consommation et modes de vie. Chroniques du CREDOC, n°18.

BORGES, J. L. (1957).- "La langue analytique de John Wilkins". *Enquêtes*, 1937-1942. NRF, Gallimard.

BOSCHER Françoise (1989).- *Opinion des Français sur les transports,* Collection des rapports, n°72, CREDOC.

BRIAN, Eric (1983).- *Le système d'enquêtes sur les Conditions de Vie et Aspirations des Français 1978-1984, phase IV. Pratiques et perceptions des transports*. - Rapport CREDOC, n°4865, vol III.

BRIAN, Eric (1984).- *Analyse des données lexicométriques. Elaboration des programmes.* Rapport CREDOC, n°4909.

BRIAN, Eric (1986).- *Techniques d'estimation et méthodes factorielles, exposé formel et application aux traitements de données lexicométriques*. Thèse de Docteur Ingénieur, Orsay.

BRISSON, L., BENZÉCRI, Jean-Pierre (1989).- "Structure de la seconde partie du Parménide". *Les cahiers de l'analyse des données*, Vol XIV, n° 3.

BRUNET, Etienne- *Hyperbase, Notice d'utilisation*, CUMFID n°17, 84 p.

CADIOT Pierre, HABERT Benoît, JACQUEMIN Christian (1992).- *Compte-rendu de la Journée d'Étude sur les Noms Composés*, École Normale Supérieure de Fontenay-St Cloud.

CENTER FOR FOOD SAFETY AND APPLIED NUTRITION (1990 a).- *LANGUAL an automated method for describing, capturing and retrieving data about food.* - Codata Conference, Columbus, Ohio, USA.

CENTER FOR FOOD SAFETY AND APPLIED NUTRITION (1990 b).- *LANGUAL Vocabulary User's Manual*, Rev.

CLERC, Laurent, DUFOUR, Ariane (1992).- *Deux analyses lexicales : les améliorations à apporter au fonctionnement de la société, et l'image du milieu professionnel.* Paris, CREDOC, Cahier de recherche, n°22.

CODOL, Jean-Paul (1969).- "Note terminologique sur l'emploi de quelques expressions concernant les activités et processus cognitifs en psychologie sociale". *Bulletin de Psychologie*, 1969, 23, pp. 63-71. (cité par Flament, 1986).

CONSTANT Patrick (1991).- *Analyse syntaxique par couche.* - Paris, Thèse de doctorat ENST.

CREDOC (Dépt. Prospective de la Consommation), Les Ateliers-ENSCI, MIND MOVERS (1992).- *Comprendre et évaluer la qualité* -. Paris, CREDOC, Cahier de recherche, n°39.

CREDOC (1993).- "Quelques notes introductives à LANGUAL".

DEARY J. (1993).- *FLAIR EUROFOODS-ENFANT Project - Langual Coding Experiment,* Ministry of Agriculture, London, U.K.

DEGENNE, A., FLAMENT, C., VERGES, P. (1973).- "Introduction à l'analyse de similitude". *Revue Française de Sociologie*, 14.- pp. 471-512.

DEGENNE, A., FLAMENT, C., VERGES, P. (1976).- *L'analyse de similitude fondée sur un indice de ressemblance.* IRIA (Classification automatique et perception par ordinateur).- pp. 249-259.

DUCROT, Oswald (1984).- *Le dire et le dit*, Paris, Ed de Minuit.

DUFLOS, Catherine, VOLATIER Jean-Luc (1991).- *Les Français et la justice, un dialogue à renouer*, Collection des rapports, n°109, CREDOC.

FEINBERG M., IRELAND-RIPERT J., FAVIER J.-C. (1991).- "LANGUAL : un langage international pour la description structurée des Aliments". *Sciences des Aliments*, 11. pp. 189-210.

FLAMENT, C., DEGENNE, A., VERGES, P. (1971).- *Similarity Analysis* -. Paris, Maison des Sciences de l'Homme, 1971.

FLAMENT, Claude (1962).- *L'analyse de similitude.* Bruxelles, Cahiers du Centre de recherche Opérationnelle, 4, 63-97.

FLAMENT, Claude (1981).- "L'analyse de similitude : une technique pour les recherches sur les représentations sociales. Commentaire sur l'article de J. P. Di Giacomo". *Cahiers de psychologie cognitive*, 1, N°4, pp. 423-428.

FLAMENT, Claude (1981).- "Sur le pluralisme méthodologique dans l'étude des représentations sociales". *Cahiers de psychologie cognitive*, 1, N°4, pp. 375-395.

FLAMENT, Claude (1989).- "Structure et dynamique des représentations sociales", Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France. pp 204-219.

FORNEL Michel de (1990).- "Sémantique du prototype et analyse de conversation", *Cahiers de linguistique française*, n° 11, Genève.

FOUCAULT, Michel (1966).- *Les mots et les choses*. Paris, Tel, Gallimard.

GOFFMAN, Erving (1987).- *Façons de parler*. - Paris, Ed de Minuit.

GREIVELDINGER Jean-François, LION Sébastien, MAISONNEUVE Christophe (1990).- *Le consommateur âgé et l'alimentation*. CREDOC, Département Prospective de la Consommation. Collection des rapports, n°92.

GRIZE, Jean-Blaise (1989).- "Logique naturelle et représentations sociales", Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France.

HAEUSLER, Laurence (1985).- *Le système d'enquêtes sur les conditions de vie et aspirations des Français. Rapport technique, vague de printemps 1985*. Rapport Crédoc n° 4927. Crédoc, Paris.

HAEUSLER, Laurence (1986).- *Le système d'enquêtes sur les conditions de vie et aspirations des Français. Rapport technique, vague de printemps 1986*. CREDOC, Collection des Rapports, n° 4946. Crédoc, Paris.

HAEUSLER, Laurence (1987).- *Analyse lexicale de réponses libres : le coût de l'électricité*, Paris, Collection des rapports, n°14.

HALBWACHS, Maurice (1925).- *Les cadres sociaux de la mémoire*. Paris, Presses Universitaires de France. Réed. Paris-La Haye, Mouton.

HARRE, Rom (1989).- "Grammaire et lexiques, vecteurs des représentations sociales", Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France.

HEARST, Marti A. (1992).- *Automatic acquisition of hyponyms from large text corpora*, Actes de COLING-92, 23-28 août 1992, pp 539-545.

HEWSTONE, Miles (1989).- "Représentations sociales et causalité", Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France. pp 252-274.

HINDE, R. A. (1971).- "La ritualisation et la communication sociale chez les singes Rhésus", Julian Huxley, ed.-. *Le comportement rituel chez l'homme et l'animal*.- Trad fr. Gallimard.- p. 73.

HOMERE (1965).- *L'Iliade*. Traduction d'Eugène Lasserre. Paris, Garnier-Flammarion.

JACQUEMIN, Christian (1990).- *Transformation des noms composés*, Paris, Thèse de doctorat Université Paris VII.

JACQUEMIN, Christian (1992).- "Activation Diffusion : A Connectionist Network for Robust Parsing" .- *Actes du Colloque E.C.A.I.*, John Wiley and Sons.- pp. 183-187.

JACQUEMIN, Christian (1992).- "Analyse des transformations des noms composés par une réseau à diffusion de l'activation".- *Actes du 5e colloque de l'ARC*, Nancy, CRIN & INRIA.- pp.45-56.

JODELET, Denise (1984).- "Représentation sociale : phénomènes, concept et théorie", Serge Moscovici (ed.) *Psychologie Sociale*, Paris, Presses Universitaires de France, 1984. pp 357-378.

JODELET, Denise (1989).- "Représentations sociales : un domaine en expansion". Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France, 1989.

JOHNSON, S. C. (1967).- "Hierarchical clustering schemes". *Psychometrica*, 32, 241-254.

KAËS, René (1984 a).- *Etayage et structuration du psychisme*. Connexions, 44, 11-48.

KAËS, René (1984 b).- "Representation and mentalisation : from the represented to group process", R. Farr and S. Moscovici (eds), *Social Representations*, Cambridge, Cambridge University Press - Paris, Maison des Sciences de l'Homme.

KAËS, René (1989).- "Psychanalyse et représentation sociale".- Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France. pp 87-114.

KRUSKAL, J. B. (1964 a).- "Multidimensional scaling by optimizing goodness of fit to a nonmetric hypothesis". *Psychometrika*, 29, 1-27.

KRUSKAL, J. B. (1964 b).- "Nonmetric multidimensional scaling, a numerical method". *Psychometrika*, 1964, 29, 115-119.

LAHLOU, Saadi (1985).- "Le nouveau produit : un concept flou".- *Consommation*, n°3.

LAHLOU, Saadi, BETBEZE, Jean-Paul, MAFFRE, Joëlle (1987 a).- *Innovation et consommation : le cas du surgelé.* Tome I : le surgelé, un marché en expansion 122 p. ; tome II : pour expliquer l'évolution des marchés alimentaires 74 p. ; tome III : le surgelé dans la consommation des ménages 158 p. + annexes 656 p. ; tome IV : le surgelé demain et après demain 1988-1995, 98 p. Rapport CREDOC.

LAHLOU, Saadi (1987 b).- *Innovation et Consommation : éléments de méthode.* Rapport CREDOC.- 100 p.

LAHLOU, Saadi (1989).- *Le comportement alimentaire des Français. Rapport au programme Aliment 2000.* CREDOC.- 118 p. + annexes.

LAHLOU, Saadi (1990).- *Eléments de formalisation pour une théorie de l'évolution des systèmes : la système-compatibilité.*- Paris, CREDOC, Cahier de recherche, n°4.

LAHLOU, Saadi, MAFFRE, Joëlle , BEAUDOUIN, Valérie (1991).- *La codification des objets complexes : réflexions théoriques et application à un corpus de 8000 produits alimentaires.*- Paris, CREDOC, Cahier de recherche, n°23.

LAHLOU, Saadi (1992 a).- *SI/ALORS : «BIEN MANGER» ? - Application d'une nouvelle méthode d'analyse des représentations sociales à un corpus constitué des associations libres de 2000 individus.*- Paris, CREDOC, Cahier de recherche du CREDOC, n°34.

LAHLOU, Saadi (1992 b).- *Modélisation et codage des aliments : théorie et application informatique pour l'Observatoire des Consommations Alimentaires*, CREDOC, octobre 1992.- 35 p. (repris dans le B.I.D. n°3, 1993).

LAHLOU, Saadi (1992 c).- "Genèse et perception de la notion de qualité", CREDOC (Département Prospective de la Consommation), Les Ateliers-ENSCI, MIND MOVERS .- *Comprendre et évaluer la qualité* -. Paris, CREDOC, Cahier de recherche, n°39.

LAHLOU, Saadi, COLLERIE DE BORELY, Aude, BEAUDOUIN, Valérie (1993).- *Où en est la consommation aujourd'hui ? : une enquête sur le consommateur français des années 90.*- Paris, CREDOC, Cahier de recherche, n°46.

LALLJEE M., ABELSON, R. P. (1983).- "The organization of explanations", M. Hewstone (ed.), *Attribution Theory : Social and functional extensions*, Oxford, Blackwell.

LE BOUEDEC, Georges (1984).- "Contribution à la méthodologie d'étude des représentations sociales". *Cahiers de Psychologie Cognitive*, vol. 4, n°3, pp. 245-272.

LEBART, Ludovic (1984).- *Sept ans de perceptions. Evolution et structure des opinions en France de 1978 à 1984. - Rapport final de la phase VII (automne 1984)*, CREDOC.

LEBART, Ludovic, SALEM, André (1988).- *Analyse statistique des données textuelles.* Préface de Christian BAUDELOT.- Paris, Dunod.

LION, Sébastien (1991).- *Constitution d'un corpus et perte d'information en analyse lexicale : méthodes et pratiques.*- Paris, CREDOC, Cahier de recherche, n°13.

LITMAN, Sonia (1993).- *Codage des plats cuisinés*, Paris, CREDOC.

MARCHOTORCHINO, F. (1989).- *Liaison analyse factorielle - Analyse relationnelle.* (I) : Dualité" Burt-Condorcet". IBM-France, F 142, Paris.

MARCHOTORCHINO, F., MICHAUD, P. (1979).- *Optimisation en analyse ordinale des données.* - Paris, Masson.

MARCHOTORCHINO, F., MICHAUD, P. (1981).- "Agrégation des similarités en classification automatique", *Revue de Statistique Appliquée*, vol. 30, n°2.

MARTIN, Robert (1992).- *Pour une logique du sens*, Paris, Presses Universitaires de France, 1992.

MAUSS, Marcel (1974).- "De quelques formes primitives de classification. Contribution à l'étude des représentations collectives". (L'année sociologique, 6, 1903) *Oeuvres de Marcel Mauss*, tome II, pp. 13-105. Paris. Les éditions de Minuit.

McCANN A. et al. (1988).- "FDA's Factored Food Vocabulary for food product description", *Journal of the American Dietetic Association*, vol.88.

MILLER George A., FELLBAUM Christiane (1991).- *Semantic Networks of English*, Cognition, 41, pp 197-229.

MINSKY, Marvin (1988).- *La société de l'esprit.*- Paris, Interéditions. -pp. 320-321.

MOSCOVICI, Serge (1984).- "Le domaine de la psychologie sociale".- Serge Moscovici (ed.) *Psychologie Sociale*, Paris, Presses Universitaires de France.- p 5-24.

MOSCOVICI, Serge (1989).- "Des représentations collectives aux représentations sociales : éléments pour une histoire", Jodelet, D. éd. *Les représentations sociales.* Paris, Presses Universitaires de France. pp. 62-86.

MOSCOVICI, Serge (1976 (1961)).- *La psychanalyse son image et son public.* -Paris, Presses Universitaires de France.

MOUNIN, Georges (1963).- *Les problèmes théoriques de la traduction.* Paris ,Gallimard.

MOUNIN, Georges (1972).- *Clefs pour la sémantique.* Paris, Seghers, 1972.

MOUNIN, Georges (1971 (1968)).- *Clefs pour la linguistique.* Paris, Seghers.

OSGOOD, C. E., SUCI, G. J. (1952).- "A measure of relation determined by both mean and difference and profile information". *Psychological Bulletin*, 49.- pp. 251-262.

PIGAMO, Frédéric (1990).- *Outils de traitement sémantique du langage naturel* - Paris, Thèse de doctorat ENST.

RASTIER, F. (1990).- *La triade sémiotique, le trivium et la sémantique linguistique, nouveaux actes sémiotiques.* - Limoges, PULIM, 9.

REINERT, Max (1983 a).- "Une méthode de classification descendante hiérarchique : application à l'analyse lexicale par contexte". - *Les cahiers de l'analyse des données*, Vol VIII, n° 2, 1983.

REINERT, Max (1983 b).- "Une méthode de classification descendante hiérarchique". *Les cahiers d'analyse des données*, 3.- pp. 187-198.

REINERT, Max (1987 a).- "Classification descendante hiérarchique et analyse lexicale par contexte : application au corpus des poésies d'Arthur Rimbaud", *Bulletin de Méthodologie Sociologique*, n°13.

REINERT, Max (1987 b).- "Un logiciel d'analyse des données textuelles : ALCESTE".- *Communication aux 5èmes Journées Internationales "Analyse de données et Informatique"*. INRIA.

REINERT, Max (1986).- "Classification descendante hiérarchique : une méthode pour le traitement des tableaux logiques de grandes dimensions", *Data Analysis and Informatics.* North-Holland, Amsterdam.- pp. 23-28.

REINERT, Max (1990).- "ALCESTE, une méthode d'analyse des données textuelles. Application au texte «Aurélia de Gérard de Nerval" *Bulletin de Méthodologie Sociologique,* 26.- pp. 25-54.

REINERT, Max (1992 a).- *La méthodologie d'analyse des données textuelles ALCESTE ; Application à l'analyse des poésies d'Arthur Rimbaud*, Conférence d'Albi.

REINERT, Max (1992 b).- *La méhodologie ALCESTE et l'analyse d'un corpus de 304 récits de cauchemars d'enfants*, Convention internationale "Ricerca qualitativa e computer nelle scienze sociali" du 30 novembre au 2 décembre 1992 à Rome.

ROSCH, E. R., MERVIS, C. B., GRAY, W., JOHNSON, D. N., BOYES-BRAEM, P. (1976).- "Basic Objects in Natural Categories".- *Cognitive Psychology*, 8.- pp. 382-439.

ROSCH, Eleanor (1975 a).- "Principles of categorization", ROSCH E. & LLOYD B. (eds), *Cognition and Categorization*, Hillsdale, (N.J.), Laurence Erlbaum Associates.- pp. 27-48.

ROSCH, Eleanor R. (1975 b).- "Universals and Specifics in Human Categorization", R. BRISLIN, S. BOSNER & W. LONNER eds. *Cross Cultural Perspectives on Learning*, Halsted, New-York.

SAGET, H (1981).- "Nature et limites de la quantification". J. PARAIN-VIAL, ed. *Les difficultés de la quantification et de la mesure*. Paris, Actes du colloque de l'Université de Dijon «Méthodologie comparée des sciences» Maloine.- pp. 265-293.

SERRES, M (1968). - *Le système de Leibniz et ses modèles mathématiques*, Paris, Presses Universitaires de France.

SERVERIN, Evelyne, BRUXELLES, Sylvie (1992).- *Le langage du dispositif des décisions judiciaires civiles : Contribution à l'étude de l'acte décisionnel*.- CERCRID-CNRS.

SHANNON, Claude E., WEAVER, Warren (1963).- *The Mathematical Theory of Communication*. - Illini Books edition. University of Illinois Press.

SPERBER, Dan (1989).- "L'étude anthropologique des représentations : problèmes et perspectives". Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France.- pp. 115-130

SPINOZA (1947).- *Ethique*, Paris, Flammarion.

STEINMETZ, Rudy (1988).- "Conceptions du corps à travers l'acte alimentaire aux XVIIe et XVIIIe siècles". *Revue d'histoire moderne et contemporaine*. Tome XXXV, CPPAP n°52558.- pp. 3-35.

TESNIÈRE, Lucien (1959).- *Eléments de syntaxe structurale*. - Paris, Editions Klincksieck.

THORPE, W. H. (1971).- "Le jeu chez les animaux". Julian Huxley, éd. *Le comportement rituel chez l'homme et l'animal*, Paris, Gallimard.- pp. 99-111.

TOURAINE, Alain (1974).- *Pour la sociologie*, Paris, Ed du Seuil.

VAN HOOF J.A.R.A.M. (1978).- "Les expressions faciales des singes catarhiniens et des anthropoïdes". Desmond Morris, ed. *L'éthologie des primates*. Editions Complexe.

WARNESSON, Isabelle (1985).- "Applied Linguistics : Optimization of Semantic Relations by Data Aggregation Techniques". - *Journal of Applied Stochastic Model and Data Analysis*, vol. 1, n°2, 1985.- pp. 121-143.

WEINBERG, Achille (1991).- "Les sciences cognitives : de nouvelles sciences de l'esprit ?", *Sciences Humaines*, n°8.- p. 28.

WHORF, Benjamen Lee (1956 a).- "On the Connection of Ideas". - in *Language, Thought and Reality*. John B. Caroll ed. The MIT Press, Cambridge, Massachussetts.- pp. 35-39.

WHORF, Benjamen Lee (1956 b).- "A Linguistic Consideration on Thinking in Primitive Communities".- *Language, Thought and Reality*. John B. Caroll ed. The MIT Press, Cambridge, Massachussetts.- pp. 65-86.

WINDISH, Uli (1989).- "Représentations sociales, sociologie et sociolinguistique". Jodelet, D. éd. *Les représentations sociales*. Paris, Presses Universitaires de France, France.- pp 169-183.

WITTGENSTEIN, Ludwig (1965 (1958)).- *Le cahier bleu*.- Paris, Gallimard, 1965.

YVON, François (1989).- "Etude sur les comportements alimentaires : les corps gras solides". Rapport CREDOC, Département Prospective de la Consommation.

YVON, François (1990).- *L'analyse lexicale appliquée à des données d'enquête : état des lieux*. - Paris, CREDOC, Cahier de recherche, n°5.

YVON, François (1990).- *Une classification empirique des produits alimentaires - Travaux complémentaires sur l'évolution des marchés alimentaires*.- Paris, CREDOC, Collection des Rapports, n °85.

ZIPF, G. K. (1974).- La psychobiologie du langage, Paris, RETZ-CEPL.

SOMMAIRE DES GRAPHIQUES

Graphique 1 : Courbes de Zipf sur les deux corpus de réponses (auto-administré, panel) 16
Graphique 2 : Nombre de mots de basse fréquence (détail de la courbe de Zipf) 16
Graphique 3 : Les évocations associées au petit déjeuner idéal (téléphone plus auto-administré) 27
Graphique 4 : Les constituants du petit déjeuner français - panel Taylor Nelson 28
Graphique 5 : Les constituants du petit déjeuner anglais - panel Taylor Nelson 30
Graphique 6 : Les constituants du petit déjeuner allemand - panel Taylor Nelson 30
Graphique 7 : Typologie des représentations et mode de questionnement 34
Graphique 8 : Les réponses des femmes selon le mode de questionnement 35
Graphique 9 : Typologie des réponses à la question sur l'effet de serre 44
Graphique 10 : Incidence de la reconnaissance des locutions sur l'analyse lexicale 46
Graphique 11 : Typologie des réponses : effet de la réduction sémantique 60
Graphique 12: Typologie sur l'ensemble du questionnaire 74
Graphique 13 : Typologie des décisions judiciaires civiles 77
Graphique 14 : Typologie des potages et soupes liquides 80
Graphique 15 : Graphe des représentations du beurre (SPAD.T) 111
Graphique 16 : Le réseau d'association du beurre 112
Graphique 17 : Extrait du réseau "beurre" 113
Graphique 18 : Typologie des représentations du beurre (ALCESTE) 117
Graphique 18 : Graphe des représentations du beurre (ALCESTE) 118
Graphique 20 : Les trois dimensions de la représentation du beurre 119
Graphique 21 : Répartition des diplômes en fonction de l'âge 126

SOMMAIRE DES TABLEAUX

Tableau 1 : Comparaison de la richesse lexicale sur les deux corpus et effet de la réduction 18
Tableau n° 2 : Les mots les plus fréquents (en auto-administré et par téléphone) 21
Tableau n° 3 : Les mots-outils les plus fréquents (en auto-administré et par téléphone) 22
Tableau 4 : Répartition des réponses suivant le sexe et le mode d'enquête 35
Tableau 5 : Tableau récapitulatif des classifications obtenues selon le mode de lemmatisation 44
Tableau 6 : Tableau récapitulatif des classifications obtenues selon le mode de réduction 60

GLOSSAIRE

antonyme :

Unité lexicale (mot, etc.) ou syntagme dont le sens est contraire (à un autre de même nature).

(Le Robert électronique).

écholalie :

Psychiatrie. Répétition automatique des paroles (ou chutes de phrases) du locuteur, observée dans certains états démentiels ou confusionnels.

(Le Robert électronique).

Par extension : désigne le fait de répéter les termes de la question dans la réponse.

énoncé :

Un énoncé : segment de discours (oral : parole, ou écrit) produit par l'énonciation d'un locuteur et dont les limites sont variables selon les critères (phonétiques; syntactiques, etc.).

(Le Robert électronique).

hapax :

Mot n'apparaissant qu'une fois dans le corpus.

homéomorphe :

chim. Se dit de corps qui présentent des analogies dans leur réseau cristallin. Par extension se dit d'objets qui présentent des analogies de forme.

homographie :

S'emploie en linguistique pour désigner des termes qui ont la même forme graphique

hyperonyme :

L'hyperonyme est "un mot (et notamment, nom) dont le sens (compréhension logique) inclut celui d'autres mots (noms)"

(Le Robert électronique).

hyperonyme :

◊ Ling. Mot (et notamment, nom) dont le sens (compréhension logique) inclut celui d'autres mots (noms). «Cheval» est l'hyperonyme de «canasson», «destrier», etc. > Générique. «Fleur» est l'hyperonyme de tous les noms de fleurs.

(Le Robert électronique).

hyperonymie :

◊ Caractère d'un hyperonyme, superordination, inclusion par un signifié du signifié d'autres mots.

(Le Robert électronique).

hyponyme :

Ling. Mot, et, notamment, nom, dont le sens (compréhension logique) est inclus dans le sens d'un autre mot. «Cheval» est un hyponyme d'«animal»; «étalon» et «destrier» sont des hyponymes de «cheval».

(Le Robert électronique)

hyponymie :

◊ Rapport de subordination (inclusion unilatérale) d'un signifié à un autre; fait, pour un mot, un nom, d'être l'hyponyme d'un autre.

(Le Robert électronique)

idiosyncrasique

2. Didact. Propre au tempérament de quelqu'un.

(Le Robert électronique)

lemmatiser :

Didact. Regrouper (les formes différentes telles qu'elles apparaissent dans un texte) sous une forme unique, de manière à définir des unités lexicales.

(Le Robert électronique)

lexème :

3. Unité lexicale en langue, intégrant la totalité des réalisations.

(Le Robert électronique)

méréonymie, méréologie :

Relation de la partie au tout

paradigme :

◊ 2. (1943). Ling. Ensemble des termes qui peuvent figurer en un point de la chaîne parlée, axe des substitutions. > Paradigmatique (série). - (S'oppose à syntagmatique). Paradigme des noms pouvant servir de sujet à un verbe, des adjectifs pouvant être épithètes d'un substantif.

(Le Robert électronique).

polysémie :

En linguistique diachronique, Caractère d'un signe conçu comme unique d'après l'étymologie* et qui s'emploie avec plusieurs sens à une même époque (par ex. mine [de crayon] et mine [d'or], tous deux issus du gallo-romain *mina, relèvent de la polysémie, alors que mine [d'or] et mine [du visage], ce dernier étant issu du breton min, relèvent de l'homonymie).

En linguistique synchronique, Caractère d'un signe qui, à fonction constante (verbe, nom, etc.) possède plusieurs signifiés à noyau commun. La polysémie ne peut se définir que par rapport à l'homonymie* et vice versa (ainsi, pour mine [de crayon] et mine [d'or], il y a homonymie et non polysémie).

(Le Robert électronique).

stématisation :

On réduit non seulement les flexions, mais aussi les suffixages.

synecdoque :

Didact. Figure de rhétorique qui consiste à prendre le plus pour le moins, la matière pour l'objet, l'espèce pour le genre, la partie pour le tout, le singulier pour le pluriel... ou inversement (ex. : les mortels pour les hommes; un fer pour une épée; l'ennemi pour les ennemis, une voile pour un navire, etc.).

(Le Robert électronique).

synonyme :

◊ 1. Ling. Se dit de mots ou d'expressions qui ont le même sens (ou une signification voisine).

(Le Robert électronique).

synonymie :

1. Relation entre deux mots ou deux expressions synonymes ; le fait linguistique que constitue l'existence de mots synonymes.

(Le Robert électronique).

troponyme :

Se dit de mots (plus particulièrement de verbes) dont l'un est l'hyponyme de l'autre d'une certaine manière.

troponymie :

Relation entre deux mots troponymes.

CAHIER DE RECHERCHE

Récemment parus :

Cross-section versus time-series income elasticities : an estimation on a pseudo-panel of Canadian surveys (1982-1990), par François Gardes, Mohamed Goaïed, Simon Langlois - Suivre le comportement automobile des ménages au cours de leur cycle de vie, par Jean-loup Madre, Caroline Gallez - Le parc automobile des ménages se diversifie, par Jean-Loup Madre, N° 43, Février 1993.

Information et organisation de la production : vers une division cognitive du travail, par Philippe Moati, El Mouhoub Mouhoud - Dynamique et mouvements longs dans la théorie évolutionniste du changement économique, par Philippe Moati - Méthodologie d'analyse de secteurs : pour une reformulation évolutionniste du "paradigme S-C-P", par Philippe Moati, N° 44, Avril 1993.

Durée du travail et conflits entre vie personnelle et vie professionnelle, par Franck Berthuit, Jean-Luc Volatier, N° 45, Avril 1993.

Où en est la consommation aujourd'hui ? Une enquête sur le consommateur français des années 90, par Saadi Lahlou, Aude Collerie de Borely, Valérie Beaudouin, N° 46, Avril 1993.

La qualité des enquêtes auprès de populations en difficulté d'insertion sociale, par Patrick Dubéchot, Michel Legros, N° 47, Avril 1993.

Président : Bernard SCHAEFER Directeur : Robert ROCHEFORT
142, rue du Chevaleret, 75013 PARIS - Tél. : (1) 40.77.85.00

CRÉDOC

Centre de recherche pour l'Étude et l'Observation des Conditions de Vie